VŠB – Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra informatiky

Centrality v sociálních sítích odvozené od degree centrality

Centrality in Social Networks Derived from the Degree Centrality

2013 Lukáš Domin

VŠB - Technická univerzita Ostrava
Fakulta elektrotechniky a informatiky
Katedra informatiky

Zadání diplomové práce

Student: **Bc. Lukáš Domin**

Studijní program: N2647 Informační a komunikační technologie

Studijní obor: 2612T025 Informatika a výpočetní technika

Téma: Centrality v sociálních sítích odvozené od degree centrality
Centrality in Social Networks Derived from the Degree Centrality

Zásady pro vypracování:

Cílem práce je seznámit se v sociálních sítích s oblastí centralit, které jsou odvozené od degree centrality. Určit výhody a nevýhody jednotlivých centralit, naimplementovat je, provést experimenty a výsledky pak přehledně vizualizovat.

1. Prostudujte a seznamte se s oblastí degree centrality, eigen centrality , Katz centrality a page ranku. Uveďte oblasti použití a rozdíly mezi těmito centralitami.
2. Vytvořte program, který předzpracuje různé druhy formátů dat na vhodný tvar.
3. Implementujte všechny výše uvedené centrality.
4. Proveďte experimenty s centralitami a jejich vhodná vizualizace.

Seznam doporučené odborné literatury:

M. E. J. Newman: Networks: An Introduction
Linton C. Freeman: Centrality in Social networks Conceptual Clarification

Formální náležitosti a rozsah diplomové práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí diplomové práce: **Mgr. Pavla Dráždilová**

Datum zadání: 16.11.2012

Datum odevzdání: 07.05.2013

VYSOKÁ ŠKOLA BÁŇSKÁ-TECHNICKÁ UNIVERZITA Fakulta elektrotechniky a informatiky OSTRAVA

doc. Dr. Ing. Eduard Sojka
vedoucí katedry

prof. RNDr. Václav Snášel, CSc.
děkan fakulty

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

V Ostravě 29. dubna 2013 .....

Tímto bych chtěl poděkovat všem, kteří mi s diplomovou prací pomohli. Především své vedoucí práce Mgr. Pavle Dráždilové, Ph.D., za spolupráci a její cenné rady.

Abstrakt

Tato práce se zabývá centralitami odvozenými od degree centrality a jejich využití při zkoumání sociálních sítí. Zaměřuje se především na degree centralitu samotnou a pak eigenvector centralitu, Katz centralitu a PageRank. Tyto centrality jsou teoreticky popsány a jejich výpočet je pak aplikován na sociální síť uživatelů Česko-Slovenské filmové databáze. Pro zisk dat a výpočty nad nimi byl napsán vlastní program. Získané informace jsou vyhodnoceny a jsou vyvozeny závěry.

Klíčová slova: Centrality, degree centralita, eigenvector centralita, Katz centralita, PageRank, sociální síť, ČSFD

Abstract

This thesis inquires into centralities derived from Degree Centrality and their application in social networks examination. It focuses mainly on Degree Centrality itself, Eigenvector Centrality, Katz Centrality and PageRank. These centralities are described in theory and their calculation is applied to Česko-Slovenská filmová databáze social network. Self-made application has been written for the purpose of data gain and computation of centralities based on these data. Gained information has been evaluated and conclusions have been deduced.

Keywords: Centralities, Degree Centrality, Eigenvecotor Centrality, Katz Centrality, PageRank, social network, ČSFD

Seznam použitých zkratek a symbolů

ČSFD	–	Česko-Slovenská filmová databáze
HTML	–	Hyper Text Markup Language
URL	–	Uniform Resource Locator
WWW	–	World Wide Web
SQL	–	Structured Query Language

Obsah

Seznam tabulek

Seznam obrázků

Seznam výpisů zdrojového kódu

1 Úvod

Zkoumání sítí je stále užitečnější a rozšířenější vědeckou disciplínou. Analyzují se systémy biologické, fyzikální, sociální a mnohé jiné. Pokud známe strukturu sítě, můžeme z ní vypočíst mnoho zajímavých ukazatelů a metrik. Sítě jsou obecně reprezentovány grafy a ty jsou pak převáděny na jiné vhodné grafové interpretace z důvodu pohodlnější práce s nimi na výpočetních systémech. Tato práce je zaměřena na zkoumání sociální sítě, tedy vazeb mezi lidmi. V případě analýzy sociálních sítí se tedy jedná o studium vztahů mezi osobami prostředky teorie grafů. V našem případě nebudeme zkoumat síť založenou na interpersonální komunikaci mezi jedinci, ale na základě podobného chování. Ukážeme, že i v takovýchto sítích můžeme využít centralit odvozených od degree centrality pro zisk zajímavých informací a získat jiný náhled na strukturu jedinců v komunitě. Konkrétně bude zkoumán vkus uživatelů veřejného internetového portálu.

Text se zabývá degree centralitou a od ní odvozenými centralitami, eigenvector centralitou, Katz centralitou a PageRankem. Nejdříve budou ustanoveny základní pojmy z teorie grafů (kapitola 2.1) a uvedené centrality budou poté teoreticky popsány (kapitoly 2.2, 2.3, 2.4, 2.5). Pro praktické zkoumání centralit byla zvolena sociální síť Česko-Slovenské filmové databáze (ČSFD), nad kterou budou tyto centrality vypočítávány. Jedná se o sociální síť fanoušků kinematografie, kde si uživatelé sdělují své názory na filmy prostřednictvím bodového hodnocení. ČSFD není zaměřena výhradně na filmy, ale také na seriály, režiséry, herce a jiné. Tato sociální síť je blíže popsána v kapitole 3. Cílem je nejen praktická aplikace výpočtu zkoumaných centralit, ale také zisk relevantních dat o vkusu uživatelů v této síti. Bude popsán proces extrakce dat (kapitola 5) a jejich zpracování (kapitola 6). V této sociální síti pak budou vypočítávány zkoumané centrality pro jednotlivé skupiny uživatelů (kapitola 7). Získané centrality budou rozebrány a z těchto výsledků budou vyvozeny závěry o dané sociální síti.

2 Centrality odvozené od degree centrality

Centrality jsou jedním z ukazatelů popisující sítě. U centralit odvozených od degree centrality se jedná o přístup ke zkoumání sociálních sítí z hlediska měření jakési důležitosti jednotivých osob v síti. Tato důležitost může mít mnoho podob a vždy záleží na konkrétní síti a aplikované metodě. Zkoumaná síť bude popsána v následujících kapitolách, ale nejdříve budou přiblíženy různé centrality a metody jejich získávání.

2.1 Základní pojmy

V této části budou uvedeny základní pojmy z teorie grafů, s nimiž budeme v textu pracovat. Bylo čerpáno z odlišných, níže uvedených zdrojů.

2.1.1 Graf

Na graf se můžeme dívat jako na grafickou reprezentaci objektů či míst a jejich vzájemného propojení na základě nějakého vztahu. Objekty mohou být například lidé a propojení může být například na základě příbuzenského vztahu. Nebo jiný příklad, křižovatky ve městě můžeme považovat za místa a ulice mezi nimi za jejich spojení. Rozdělujeme dva základní druhy grafu, a to podle toho, zda-li záleží na orientaci hran.

Definice 2.1 *Neorientovaný graf je dvojice* $G = \langle V, E \rangle$*, kde* V *je neprázdná množina tzv. vrcholů (někdy také uzlů) a* $\mathrm{E} \subseteq \{\{u, v\} \mid u, v \in V\}$ *je množina dvouprvkových či jednoprvkových množin vrcholů, tzv. neorientovaných hran [11].*

Na obrázku 1 je ukázka neorientovaného grafu $G_1 = \langle V, E \rangle$, kde $\mathrm{V} = \{U, R, X, Y, Z, S, T, W\}$ a $\mathrm{E} = \{a, b, c, d, e, f, g, h, i, j, k, l, m\}$.

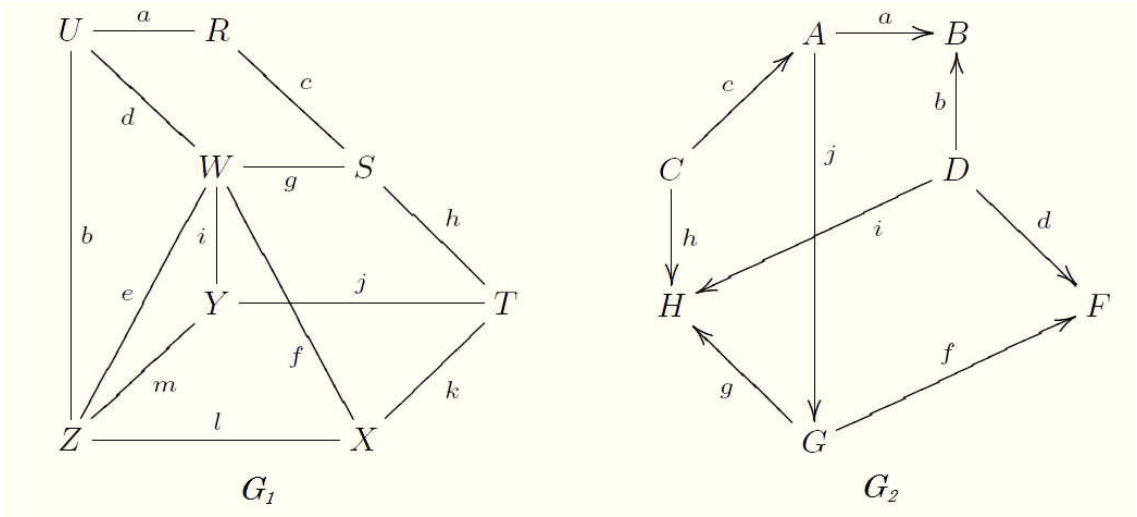
Definice 2.2 *Orientovaný graf je dvojice* $G = \langle V, E \rangle$*, kde* V *je neprázdná množina tzv. vrcholů (uzlů) a* $\mathrm{E} \subseteq V \times V$ *je množina uspořádaných dvojic vrcholů, tzv. orientovaných hran [11].*

Na obrázku 1 je ukázka orientovaného grafu $G_2 = \langle V, E \rangle$, kde $\mathrm{V} = \{A, B, C, D, E, F, G, H\}$ a $\mathrm{E} = \{a, b, c, d, e, f, g, h, i, j\}$.

Jednotlivým vrcholům, ale i hranám grafu je možno přiřadit číselné ohodnocení, a tím získáme ohodnocený graf. Ohodnocení vrcholů reprezentujících objekty může vyjadřovat například jejich velikost. Naopak ohodnocení hran může vyjadřovat třeba vzdálenost mezi těmito objekty.

Definice 2.3 *Buď G graf s množinou vrcholů V a množinou hran E. Nechť* $f : V \to R$ *a* $g : E \to R$ *jsou zobrazení. Pak f se nazývá vrcholovým ohodnocením grafu G, g se nazývá hranovým ohodnocením grafu G. Dvojice (G,f), resp. (G,g) se nazývá vrcholově, respektive hranově ohodnocený graf [12].*

V orientovaném grafu rozlišujeme vůči danému vrcholu dva druhy hran. Buď orientovaná hrana z vrcholu vychází a ukazuje tedy na jiný vrchol nebo naopak do vrcholu vstupuje. Příkadem může být graf sítě citací, kdy je hrana znázorňující citaci u člověka



Obrázek 1: Neorientovaný graf G_1 a orientovaný graf G_2

citujícího považována za výstupní, tedy out-degree a u člověka citovaného vstupní, tedy in-degree.

Definice 2.4 *Out-degree vrcholu v, značíme $deg^+(v)$, je počet hran vycházejících z v a in-degree vrcholu v, značíme $deg^-(v)$, je počet hran vstupujících do v. Pro každý graf G pak platí:*

$$\sum_{v \in V(G)} deg^+(v) = |E(G)| = \sum_{v \in V(G)} deg^-(v), \tag{1}$$

kde $V(G)$ jsou vrcholy grafu G a $E(G)$ jsou hrany grafu G.

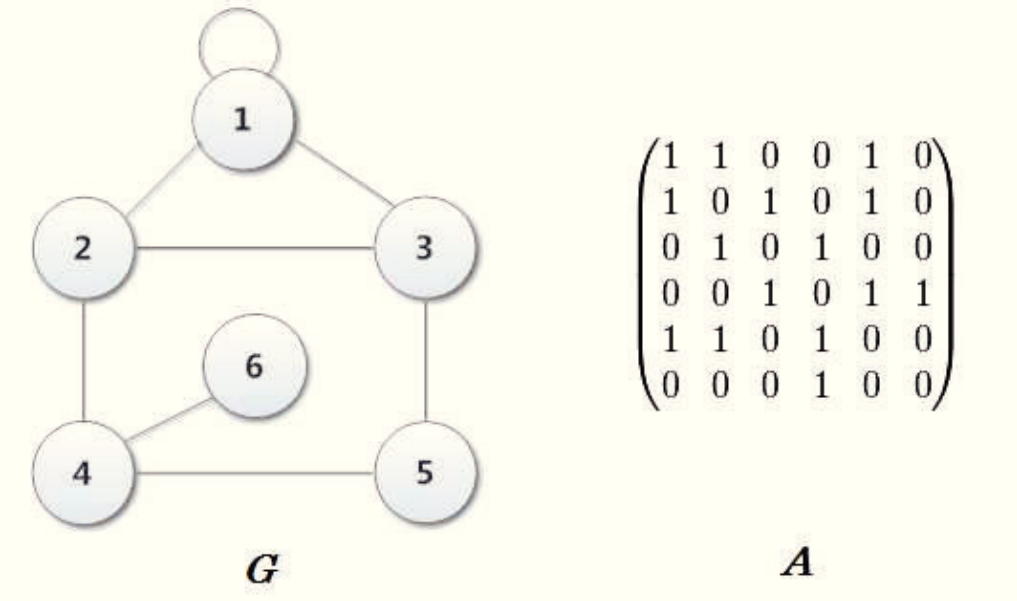
Jednou z možných reprezentací grafu je matice sousednosti (anglicky Adjacency matrix). Jedná se o čtvercovou matici o velikosti n × n, kde n je počet vrcholů v grafu. Hodnoty v této matici A_{ij} udávají vztah mezi uzly na pozicích *i* a *j*. Hodnota na místě a_{ij} udává počet hran vedoucích z vrcholu *i* do vrcholu *j*. Pokud je v matici na tomto místě nula, vrcholy nejsou propojeny. Případně mohou hodnoty představovat sílu (váhu) vazby mezi dvěmi vrcholy. Pro neorientované grafy jsou matice sousednosti diagonálně symetrické. Na obrázku 2 je ukázka jednoduchého neorientovaného grafu *G* a jeho matice sousednosti *A*.

Pokud matice obsahuje váhy hrany, nazýváme tuto matici *vážená matice sousednosti*.

Definice 2.5 *Vážená matice sousednosti ohodnoceného orientovaného grafu (G,w) s vrcholy v_1,...,v_n je matice W(G) = (w_{ij}), kde*

$$w_{ij} = \begin{cases} w(v_i v_j) & \text{pokud } e(v_i v_j) \in E(G) \\ 0 & \text{jinak,} \end{cases} \tag{2}$$

pro i, j = 1,...,n [10].



$$\begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

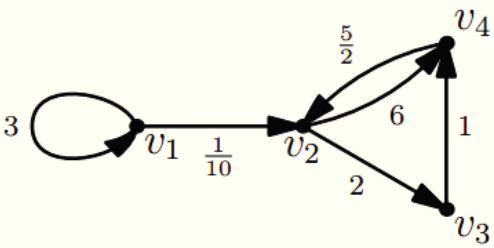
Obrázek 2: Graf G a odpovídající matice sousednosti A

Na obrázku 3 je pro ilustraci jiný, tentokrát ohodnocený orientovaný graf G a jeho vážená matice sousednosti W.

2.2 Degree centralita

Tato centralita je nejzákladnější a nejjednodušší z představených. Její velikost je závislá na počtu hran (vazeb) jednoho konkrétního vrcholu. Tedy s kolika dalšími vrcholy grafu je onen vrchol propojen. Degree centralita C_D vrcholu i v síti G se vypočte jako

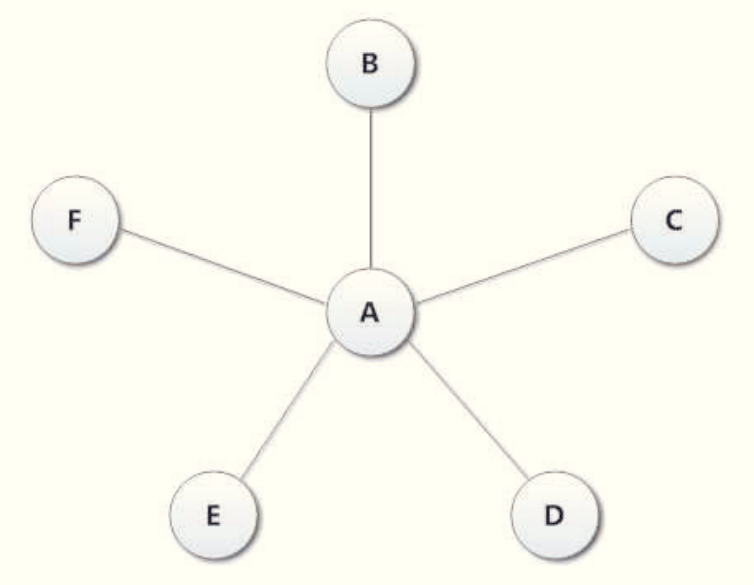
$$C_D(i) = \frac{d_i}{n-1}, \tag{3}$$



$$\begin{bmatrix} 3 & \frac{1}{10} & 0 & 0 \\ 0 & 0 & 2 & 6 \\ 0 & 0 & 0 & 1 \\ 0 & \frac{5}{2} & 0 & 0 \end{bmatrix}$$

W

Obrázek 3: Graf G a odpovídající vážená matice sousednosti W



Obrázek 4: Graf jednoduché sítě

kde d_i je počet sousedů vrcholu i (počet hran) a n je celkový počet vrcholů v síti [3]. Je zřejmé, že $0 \leq C_D(i) \leq 1$.

V orientovaných sítích můžeme rozlišovat hrany vstupní *(in-degree)* a výstupní *(out-degree)*. Například pokud uživatel A pošle zprávu uživateli B, vytvoří se mezi vrcholy A a B reprezentujícími tyto uživatele vazba. Z pohledu uživatele A, který zprávu zaslal, se jedná o out-degree hranu a z pohledu přijímatele B o in-degree. Jestli budeme tuto orientaci rozlišovat, je už na nás. Ačkoliv se jedná o velmi jednoduchou centralitu, může pro nás mít často velkou vypovídající hodnotu. Třeba v sociálních sítích, ve kterých dochází ke komunikaci mezi osobami, můžeme předpokládat, že čím má osoba více vazeb na jiné (více zaslaných příp. přijatých zpráv), tím je v síti nějak významnější. Ačkoliv někdy mohou být takovéto závěry zavádějící. Skutečnost, že je někdo upovídaný, ještě nutně nemusí znamenat, že je v komunitě nějak významněji důležitý. Každopádně vzhledem k jednoduchosti degree centrality je její význam veliký, jelikož nám dokáže velmi rychle popsat jistý aspekt sítě. Této metrice se v sociálních sítích také přezdívá *hledání celebrit* [4]. Tento název vyjadřuje právě onu skutečnost, že vrcholy s nejvyšší degree centralitou nemusí být nutně v pravém slova smyslu nejdůležitější v síti, ale pouze jen z jistého pohledu zajímavé. Na následujícím obrázku 4 je znázorněna velmi jednoduchá síť o šesti vrcholech.

Pokud pomineme skutečnost, že v reálných sociálních sítích jsou takovéto hvězdicové konfigurace velmi vzácné a vrcholy si představíme jako osoby a hrany jako komunikaci mezi nimi, snadno odhadneme, že v této síti je uživatel A výjimečný. Veškerá komunikace směřuje k němu a tedy prochází skrze něj. Tato pozice by jej tedy měla činit v síti důležitým. Ve skutečnosti ale záleží na povaze vazeb, čili vztahů mezi osobami. Vztahy mohou být velmi různorodé a jejich výpovídající hodnota může být větší i menší. Sami z vlastní

zkušenosti ze společnosti víme, že nezřídka mívají největší vliv osoby zdánlivě nenápadné, cíleně zůstávající v pozadí.

2.2.1 Freemanova modifikace

Linton Freeman vyvinul metriku centrality, kdy nezáleží pouze na stupni centrality vrcholů, ale také na tzv. centralizovanosti celé sítě. Freemanův přístup vyjadřuje stupeň variability stupně centrality vrcholů v síti jako procentuální poměr k variabilitě v případě dokonalé hvězdicové konfigurace sítě stejné velikosti [6]. Příklad hvězdicové sítě je na obrázku 4. Dokonalá hvězdicová síť o *n* vrcholech má právě jeden vrchol se stupněm centrality rovným *n-1* a *n-1* vrcholů se stupněm centrality rovným jedné. Centralizovanost neorientované sítě *G* se pak vypočte jako:

$$C_G = \frac{n\Delta_G - m}{(n-1)(n-2)}, \tag{4}$$

kde *n* je počet vrcholů v síti, *m* je počet hran v síti a Δ_G je nejvyšší stupeň degree centrality ze všech vrcholů sítě. V případě orientovaných sítí je vzorec obdobný:

$$C_G^- = \frac{n\Delta_G^- - m}{(n-1)(n-2)}, \tag{5}$$

pro in-degree a

$$C_G^+ = \frac{n\Delta_G^+ - m}{(n-1)(n-2)}, \tag{6}$$

pro out-degree. Tento údaj se uvádí v procentech. Pokud máme síť o následujících parametrech, n = 11, m = 49, Δ_G^- = 9, Δ_G^+ = 8, budou výpočty:

$$C_G^- = \frac{11 \times 9 - 49}{10 \times 9} = \frac{50}{90} \doteq 55,6\% \tag{7}$$

pro in-degree a

$$C_G^+ = \frac{11 \times 8 - 49}{10 \times 9} = \frac{39}{90} \doteq 43,3\% \tag{8}$$

pro out-degree. Tato metrika je použita například ve vědeckém článku [7] zaměřeném na studium vztahů mezi imunologickými organizacemi v Baltimoru.

2.2.2 Bonacicho modifikace

Dobrým příkladem výjimky v tradičním pojetí centrality jsou výměnné sítě. Jde o sítě, kde dochází k výměnnému obchodu mezi osobami, kus za kus. Cook et al. (1983) ukázal, že u těchto sítí se síla nerovná centralitě [5]. V těchto sítích dochází ke smlouvání a pro jedince je naopak výhodné být spojen pouze s bezmocnými, kteří nemají jinou možnost. Pokud by byl naopak spojen s mnoha silnými jedinci, jeho síla možnosti smlouvání by se snížila. Pro tyto sítě navrhl Bonacich následující modifikaci degree centrality, která je univerzálnější.

$$C_B(i) = \sum_j (\alpha + \beta C_B(j)) A_{ij}, \tag{9}$$

kde $C_B(i)$ je centralita daného vrcholu, $C_B(j)$ jsou centrality okolních vrcholů, A_{ij} je matice sousednosti a parametry α a β jsou vhodně zvolené konstanty. Při zvolení $\beta < 0$ je centralita vrcholu, mající sousedy s vrcholy s velkou centralitou, snižována. V tomto případě tedy nezáleží pouze na tom, kolik má vrchol sousedů, ale také na tom, jací jsou tito sousedi, což je základem všech následujících centralit. Vzhledem k tomu, že centralita vrcholů závisí na okolních vrcholech, používají se zde iterační algoritmy, kdy se snažíme správnou centralitu stále přesněji vypočítávat.

2.3 Eigenvector centralita

Přirozeným rozšířením jednoduché degree centrality je eigenvector centralita. Jedná se o rozšíření tzv. horizontu pozorovatelnosti, tedy schopnosti vidět v síti na více úrovní [4]. Nyní se nebere v úvahu pouze to, kolik vazeb vrchol má, ale také kolik vazeb mají sousední vrcholy. Tomuto způsobu se v sociálních sítích také říká hledání šedé eminence, jelikož se zde projeví i osoby, které sice nemají mnoho vazeb, ale mají vazby na jiné důležité lidi. Takovéto osoby by pro nás byly v případě jednoduché degree centrality vpodstatě neviditelné. Degree centralitu si tedy můžeme představit jako případ eigenvector centrality, kdy vrchol ohodnocujeme právě jedním bodem centrality za každého souseda. Tím, že nyní dělíme sousedy na více a méně důležité, ohodnocujeme různým počtem bodů. Eigenvector centralita přiřazuje každému vrcholu skóre úměrné součtu skóre jeho sousedů. Eigenvector centralita je vpodstatě rekurzivní verze degree centrality [4].

Data: Graf $G(v, e)$
Result: Eigenvector centrality $C_E(i)$ pro vrcholy $v(i)$
Všem vrcholům v síti přiřadíme skóre centrality rovno jedné ($C_E(i)$ = 1 pro všechna *i*)
while $C_E(i)$ *předchozí iterace* $\dot{\neq}$ $C_E(i)$ *akt. iterace* & *akt. iterace* $\neq$ *maximální iterace* **do**

1. Přepočítáme skóre všech vrcholů jako vážený součet centralit všech vrcholů v sousedství vrcholu. Sousednost a případnou sílu vazby určíme z matice sousednosti A.

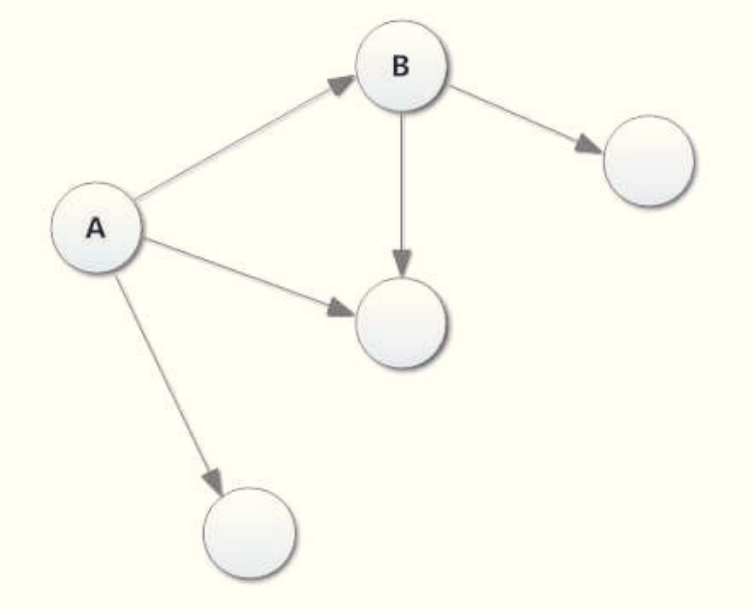
$$C_E(i) = \sum_{j \neq i} A_{ij} C_E(j), \tag{10}$$

2. Normalizujeme $C_E(i)$ podělením každé hodnoty hodnotou největšího vlastního čísla matice.

end

Algoritmus 1: Výpočet eigenvector centrality

Eigenvector centralita má tedy tu vlastnost, že může být vysoká, buď protože má vrchol mnoho hran, nebo protože má důležité sousedy nebo obojí. Osoba v sociální síti



Obrázek 5: Část orientované sítě

může být díky tomuto důležitá nejen tím, že zná mnoho jiných osob, ale také tím, že zná pár osob na vysokých místech [1]. Teoreticky je možné počítat eigenvector centrality jak pro orientované, tak pro neorientované sítě. U orientovaných je to však komplikovanější. V takovýchto sítích je matice sousednosti logicky typicky nesymetrická. Máme tedy dvě množiny vlastních vektorů - levé a pravé. Nabízí se otázka, které z nich využít. Vzhledem k tomu, že v orientovaných sítích jsou pro nás většinou důležitější hrany které míří k nám, než ty které z nás vycházejí, je obvykle správnou volbou použít pravé vlastní vektory. Lze si to jednoduše ilustrovat na příkladu World Wide Webu, kdy jsou pro určení důležitosti stránky mnohem důležitější odkazy ukazující na ni, než odkazy ukazující z ní. Toto ovšem není jediný problém u orientovaných sítí. Zaměřme se na vrchol A v obrázku 5.

Tento vrchol je spojen se sousedy pouze výstupními hranami a žádnými vstupními. Takovýto vrchol bude zákonitě mít nulovou eigenvector centralitu. To samo o sobě ještě není problém, prostě vrchol považujeme za nedůležitý. Když se ale nyní podíváme na vrchol B, tak vidíme, že má pouze jednu jedinou vstupní hranu a to právě z A. Proto má i vrchol B centralitu nula. Pokud se bude situace opakovat do více úrovní a progrese skončí u vrcholu s nulovou in-degree centralitou, je výpočet zbytečný, jelikož celková hodnota centrality bude stejně nulová. Příklad takového stavu je síť citací. V takovéto acyklické síti mají všechny vrcholy ohodnocení nula, jelikož se zde předává nulové ohodnocení celým grafem. Pro sítě toho typu je tedy eigenvector centralita nepoužitelná. Variantou eigenvector centrality řešící tyto problémy je Katz centralita, která bude probrána v následující kapitole.

2.4 Katz centralita

Řešením výše uvedených problémů je jednoduché přičtení malé konstantní hodnoty centrality ke každému vrcholu v síti, nezávisle na jeho pozici v síti či centralitě sousedů [1]. Vzorec 10 bude pro výpočet Katz centrality $C_K(i)$ upraven následovně:

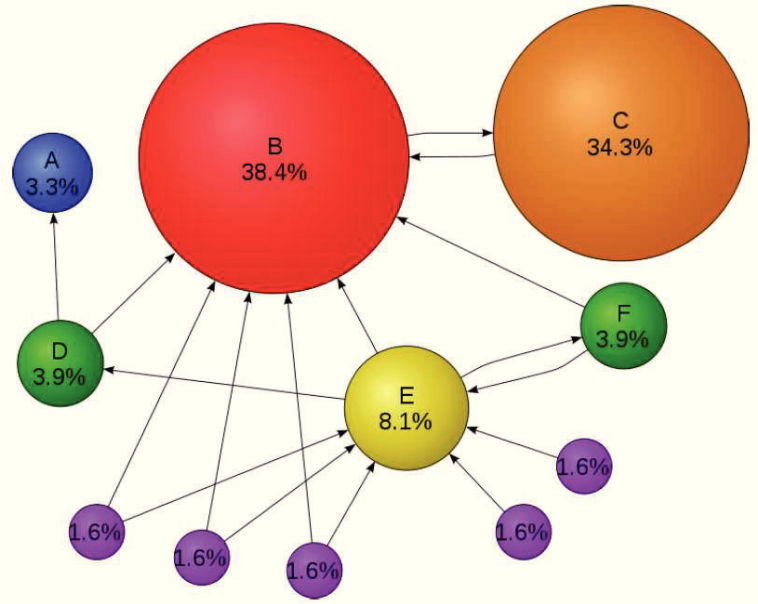
$$C_K(i) = \alpha \sum_{j \in N} A_{ij} C_K(j) + \beta, \tag{11}$$

kde α a β jsou kladné konstanty. Přičtením konstanty získají nenulovou centralitu i ty vrcholy, které by jinak byly v případě eigenvector centrality ohodnoceny nulou. To znamená, že každá vazba bude nyní přidávat alespoň nějakou malou váhu. Vzhledem k tomu, že obvykle nám na absolutní velikosti centrality nezáleží, ale spíše pouze na tom, které vrcholy ji mají velkou a které malou, je velikost konstanty β nedůležitá. Z praktických důvodů se konstanta β většinou dává rovna jedné. Konstanta α slouží k udržování rovnováhy mezi výrazem pro výpočet eigenvector hodnoty (vzorec 10) a konstantou β. Při určování velikosti této konstanty musíme mít na mysli, že pokud zvolíme α příliš malé ($\alpha \rightarrow$0), poté se ve vzorci 11 projeví pouze konstanta β a všechny vrcholy budou mít stejnou centralitu. Aby centralita konvergovala, nesmí být α naopak ani příliš velká. Musíme vždy vybírat hodnotu menší než $1/\kappa_1$, kde κ_1 je největší hodnota ve vlastním vektoru A [1]. Krom těchto omezení je ale určení přiměřené hodnoty α na našem odhadu. Ačkoliv byla jako hlavní důvod zavedení Katz centrality uvedena její schopnost vypořádat se s problémy u orientovaných sítí, je samozřejmě možné využívat Katz centralitu i u neorientovaných sítí. Katz centralitu je možno ještě vylepšit úpravou parametru β. Místo konstantní hodnoty, stejné pro všechny vrcholy, může β nabývat různých hodnot pro různé vrcholy dle nějaké metriky nesouvisející se samotnou sítí. Například věk osob v sociální síti. Tím je možno elegantně zvyšovat důležitost určitých individuí. Vzorec výpočtu by pak vypadal následovně:

$$C_K(i) = \alpha \sum_{j \in N} A_{ij} C_K(j) + \beta_i \tag{12}$$

2.5 PageRank

Ačkoliv byl PageRank původně vyvinut pro indexování webových stránek, je možno jej aplikovat také na (sociální) sítě [4]. PageRank je pojmenován po zakladateli společnosti Google Larrym Pageovi a tento název je obchodní známkou společnosti Google a postup PageRanku je patentován. Patent ovšem patří univerzitě Stanford a Google má "pouze" výhradní právo pro jeho používání. PageRank můžeme opět brát jako jakési rozšíření předchozí Katz centrality. Katz centralita totiž má jistou pro nás potencionálně nežádoucí vlastnost. A totiž tu, že pokud vrchol s vysokou Katz centralitou ukazuje na mnoho jiných vrcholů, všechny tyto vrcholy také dostanou vysokou centralitu. To může být zkreslující a tedy nežádoucí. Pokud si představíme nějakou velmi důležitou webovou stránku, odkazující na milión jiných stránek, všechny obdrží tuto velikou hodnotu. Takto ony stránky získají neoprávněně na důležitosti. Úprava Katz centrality spočívá v tom, že centralita, kterou vrchol získá od sousedních vrcholů, je úměrná jejich centralitě podělené



Obrázek 6: Ilustrace PageRanku

jejich out-degree. Poté vrcholy ukazující na mnoho jiných předávájí ostatním pouze malou část jejich centrality, přestože je jejich centralita vysoká [1]. Postup výpočtu PageRanku $C_P(i)$ tedy můžeme zapsat jako:

$$C_P(i) = \alpha \sum_{j \in N} A_{ij} \frac{C_P(j)}{k_j^+} + \beta \tag{13}$$

k_j^+ značí out-degree, tedy na kolik jiných vrcholů onen vrchol odkazuje. Je zřejmé, že pokud se v grafu objeví vrcholy s out-degree $k_j^+ = 0$, máme problém. V tom případě bychom totiž dělili nulu nulou ($A_{ij} = 0$ pro všechna i). To lze však jednoduše obejít. Pokud vrchol nemá žádné výchozí hrany, stejně nemůže nijak přispívat do centralit ostatních. Pro tyto vrcholy tedy můžeme uměle nastavit třeba $k_j^+ = 1$ a problém je vyřešen [1]. Pro ilustraci PageRanku se často používá obrázek 6 získaný z [4].

Na něm lze dobře vidět jak PageRank funguje. Můžeme na něm například vidět, že vrchol C získal vysoké ohodnocení, přestože na něj ukazuje pouze vrchol B. Vrchol B má ovšem jednak velmi vysoké ohodnocení, jelikož na něj odkazuje mnoho jiných, ale také sám odkazuje jen a pouze na vrchol C a tím mu předává velkou centralitu, jelikož se nemusí dělit s dalšími vrcholy. Parametr α se pro PageRank nastavuje téměř výhradně na hodnotu $\alpha = 0,85$ [1]. Nastavení parametru β se různí. Nejčastěji se používají hodnoty dle vzorců 14, 15, 16. Ve vzorci 16 N značí počet vrcholů v síti.

$$\beta = 1 \tag{14}$$

$$\beta = 1 - \alpha \tag{15}$$

$$\beta = \frac{1-\alpha}{N} \tag{16}$$

Existují také vážené varianty PageRanku. V orientované síti je možno vypočítávat PageRank podle 7, kdy se váha (také nazývaná polularita) vrcholu vypočte na základě poměru příchozích a odchozích hran a je vyjádřena v parametrech $W^{in}_{v,u}$ a $W^{out}_{v,u}$. $W^{in}_{v,u}$ je váha hrany $e(v,u)$ a je vypočtena na základě počtu vstupních hran vrcholu u a počtu výstupních hran všech sousedů vrcholu v.

$$W^{in}_{v,u} = \frac{I_u}{\sum\limits_{p \in R(v)} I_p} \tag{17}$$

kde I_u a I_p reprezentují počet vstupních hran vrcholů u a p. $R(v)$ značí množinu sousedních vrcholů vrcholu v. $W^{out}_{v,u}$ je váha hrany $e(v,u)$, vypočtena na základě počtu výstupních hran vrcholu u a počtu výstupních hran všech sousedů vrcholu v.

$$W^{out}_{v,u} = \frac{O_u}{\sum\limits_{p \in R(v)} O_p} \tag{18}$$

kde O_u a O_p reprezentují počet výstupních hran vrcholů u a p. $R(v)$ značí množinu sousedních vrcholů vrcholu v. Poté se vážený PageRank vrcholu u vypočte jako:

$$C^w_P(u) = \alpha \sum_{v \in N} C^w_P(v) W^{in}_{v,u} W^{out}_{v,u} + \beta \tag{19}$$

Ve vážené síti, kdy jsou jednotlivým hranám přiřazeny váhy, je možno vypočítávat vážený PageRank dle vzorce:

$$C^w_P(i) = \sum_i \alpha C^w_P(j) \frac{A_{ij}}{w_{sum}(j)} + R, \tag{20}$$

kde A_{ij} je matice sousednosti obsahující váhy hran, $w_{sum}(j)$ je součet vah všech hran vrcholu j a R se vypočte podle vzorce:

$$R = \sum_i \beta \frac{C^w_P(i)}{N}, \tag{21}$$

kde N značí počet vrcholů v síti a $C^w_P(i)$ je vypočtený PageRank vrcholu i z předchozí iterace. Tento výpočet se používá například v programu Gephi [15].

PageRank je iterativní proces, jinak známý pod anglickým termínem *anytime algorithm* [4]. Ten vyjadřuje skutečnost, že vrátí kdykoliv výsledek, ale čím více času (a tedy iterací) se mu dá, tím bude výsledek přesnější. Nejdříve dává jen hrubé výsledky, ty se ale s narůstajícím počtem iterací budou zlepšovat, dokud nedokonvergují do bodu stability nebo dokud uživatel sám výpočet nezastaví.

2.5.1 HITS

Hyperlink-Induced Topic Search (HITS) je předchůdcem PageRanku. Nejedná se však vyloženě o druh centrality jako předchozí. Tento algoritmus pro analýzu vazeb (odkazů)

ve WWW síti byl vyvinut Jonem Kleinbergem. HITS operuje s dvěma základními pojmy - huby a autority. Tyto pojmy vychází z původního, dnes již zastaralého pojetí internetu, kdy jisté velké internetové stránky, nazývané huby, sloužily jako rejstříky, jež obsahovaly velké množství neautoritativních informací, které vedly k menším stránkám, zvaných autority, kde již byly autoritativní informace. Dobrý hub byl tedy stránka odkazující na mnoho jiných a dobrá autorita stránka odkazována mnoha huby. Každý vrchol v takovéto síti tedy obsahoval dva základní parametry - autoritu (hodnota obsahu) a hodnotu hubu (hodnota spojení k jiným vrcholům) [8].

Nechť a_p reprezentuje autoritu a h_p reprezentuje hodnotu hubu. $B(p)$ pak reprezentuje množinu stránek odkazujících na stránku p a $I(p)$ množinu odkazovaných stránek ze stránky p. Pak se autorita a hodnota hubu stránky p vypočte následovně:

Všechna a_p a h_p inicializujeme na počáteční hodnotu rovnou jedné.
while *iterace* $\neq$ *maximální iterace* **do**

1. Přepočítáme skóre autorit dle:

$$a_p = \sum_{q \in B(p)} h_q \tag{22}$$

2. Přepočítáme skóre hubů dle:

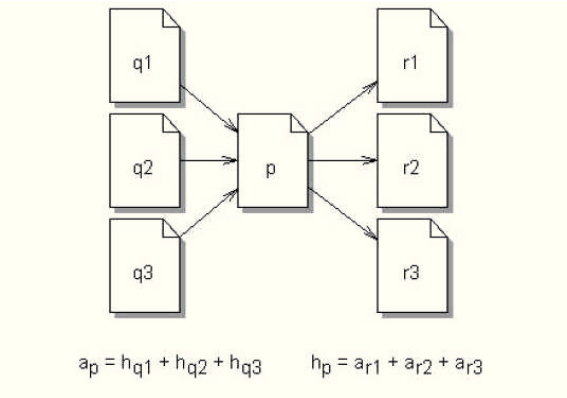
$$h_p = \sum_{q \in I(p)} a_q \tag{23}$$

3. Hodnoty skóre autorit i hubů normalizujeme.

end

Algoritmus 2: Výpočet HITS

Hodnoty se normalizují vydělením každé a_p druhou odmocninou sumy druhých mocnin všech a_p a vydělením každé h_p druhou odmocninou sumy druhých mocnin všech h_p. Na obrázku 7 je uvedeno grafické znázornění výpočtu [9]. Tento algoritmus se však již ve vyhledávačích na internetu téměř vůbec nevyužívá.



Obrázek 7: Příklad výpočtu HITS

3 Sociální síť Česko-Slovenská filmová databáze

V této kapitole bude popsán portál Česko-Slovenské filmové databáze, zkráceně ČSFD, ze kterého byla získána data. Bude brán zřetel především na ty aspekty portálu, které jsou pro nás důležité z důvodu následujícího zkoumání sociální sítě. Vazby v této síti netvoří přímo sociální interakce, ale jsou vytvořeny na základě společného chování - shodného hodnocení filmů, výběru shodných oblíbených seriálů, režisérů, skladatelů atd. V síti budou vypočteny výše uvedené centrality pro jednotlivé uživatele.

3.1 Zaměření ČSFD

Internetový portál ČSFD je rozsáhlou databází filmů, seriálů, televizních pořadů a jejich tvůrců. Nachází se na adrese www.csfd.cz a jedná se o projekt slovenského filmového nadšence Martina Pomothyo. Databáze je neustále aktualizována a obsahuje takřka kompletní seznam všech kinematografických děl, které kdy vznikly. ČSFD ovšem není pouhou databází, ale především velkou sociální sítí filmových fanoušků. Po zaregistrování nabízí možnost hodnotit filmy a seriály a také je okomentovat svými minirecenzemi. Na základě těchto hodnocení jsou vypočítávány statistické ukazatele vyjadřující oblíbenost jednotlivých děl. Hodnocení všech uživatelů jsou veřejně viditelné a uživatelé tedy mohou vzájemě porovnávat svůj vkus, diskutovat o filmech, doporučovat či naopak odrazovat od shlédnutí.

3.2 Hodnocení filmů

Každý zaregistrovaný uživatel má možnost hodnotit jak filmy, tak také seriály a televizní pořady. Hodnocení je na stupnici 0 až 5 hvězdiček. Pět hvězdiček je nejpozitivnější možné ohodnocení, naopak 0 hvězdiček (označováno jako *odpad!*) je nejnegativnější ohodnocení. Pokud má uživatel ohodnoceno alespoň 200 titulů, projevují se jeho hodnocení do celkových statistik. Na základě všech ohodnocení titulu je vypočten procentuální průměr vyjadřující kvalitu onoho titulu. Díla je tedy možno řadit podle celkového ohodnocení a uživatelé vidí, která jsou považována za kvalitní, a která nikoliv. Na základě tohoto hodnocení se mohou potencionální diváci rozhodnout, jestli mu věnují svůj čas a shlédnou ho.

3.3 Kategorie *oblíbené*

Kromě hodnocení hvězdičkami nabízí ČSFD také různé kategorie zvané *oblíbené*. Jedná se o sedm následujících kategorií:

- Filmy
- Seriály
- Pořady
- Herci

- Herečky
- Režiséři
- Skladatelé

Do každé kategorie je možno přiřadit až deset titulů či tvůrců. Tím si uživatel vytvoří svůj osobní žebříček TOP 10 oblíbených filmů, seriálů, herců atd. Zařazení filmu mezi oblíbené nemá vliv na jeho hodnocení. Jsou však generovány žebříčky nejoblíbenějších titulů a tvůrců. Například čím více lidí má určitého režiséra zařazeno mezi oblíbenými, na tím vyšší pozici je v žebříčku oblíbených režisérů.

3.4 Filmotéka

Uživatelé mají také možnost si na tomto portálu evidovat svou filmotéku. Mohou si zde do databáze zaznačit jaké filmy vlastní, na jakém médiu, v jaké kvalitě, v jakých jazycích a také s jakými titulky. Tyto filmotéky jsou také volně dostupné a uživatelé tedy navzájem vidí, kdo co vlastní. Filmotéka je tedy dalším propojením uživatelů v síti na základě stejných vlastněných titulů.

3.5 Filmy

Filmy, seriály a televizní pořady mají na svých profilových stránkách uvedeno mnoho údajů. Mezi pro nás nejzajímavější patří následující:

- Žánr
- Režisér
- Scénarista
- Skladatel
- Herci
- Rok vzniku
- Země původu

Na základě těchto údajů je možno filmy třídit do různých kategorií a v návaznosti na uživatele, kteří jsou s nimi spjati, je možno získat zajímavé informace o vkusu uživatelů v této síti.

4 Zkoumání sítě ČSFD

Pro náš cíl zkoumání jsou na portálu ČSFD důležité především profilové stránky uživatelů, kde jsou uvedeny jejich hodnocení filmů a jejich oblíbené filmy, seriály, režiséři atd. Je také možné získat podrobnější informace o těchto dílech a autorech. Na základě těchto údajů může být vytvořena sít ČSFD propojující uživatele dle jejich společného vkusu v této oblasti. Je možno vytvořit hned několik druhů sítí na základě toho, podle kterých kritérií budeme konstruovat vazby mezi uživateli. ČSFD nabízí velké množství dat a je na experimentátorovi, jaké konkrétní oblasti ho zajímají, a tedy na základě jakých údajů vazby v síti vytvářet. Konstrukcí různých sítí se budeme zabývat především v kapitolách 6.2 a 7.2.

4.1 Cíle zkoumání

Musíme si zákonitě položit otázku: Jaké informace získáme zkoumáním této sítě? Konkrétně výpočtem výše uvedených centralit. Nejdříve se musíme podívat jakou síť to vlastně před sebou máme a co v ní jednotlivé vazby představují. Figurují v ní uživatelé, tedy lidé, kteří mají mezi sebou jisté vztahy - vazby. Jedná se tedy určitě o sociální síť. Ovšem ne typickou, jakou si obvykle pod tímto pojmem obvykle představujeme. Porovnáme si ji s nejtypičtější sociální sítí - internetovým diskuzním fórem. O tom určitě můžeme říci, že představuje obvyklou sociální síť, na kterou se různé studie (včetně právě počítání centralit) zaměřují. Tyto sítě si jsou do jisté míry podobné. V obou máme uživatele, kteří do sítě přispívají tím, že mezi nimi probíhají jisté interakce. Základní rozdíl je však ve vazbách mezi nimi a tím, co vyjadřují. Na fórech dochází ke komunikaci. Ať už formou veřejnou - odpovědí na příspěvek jiného uživatele, tak formou soukromých zpráv. Každou takovouto reakcí se vytváří vazba mezi právě dvěma uživateli. Tato komunikace u nich v jistém slova smyslu zvyšuje sociální důležitost. Obvykle dochází k výměně názorů, z nichž benefitují obě strany. Fóra jsou obvykle zaměřena na nějaká témata (programování, hry, automobily, divadlo atd.) a my tedy můžeme předpokládat, že komunikací dochází ke zvyšovaní odbornosti patricipovaných osob v této oblasti.

Jaké vazby však vznikají v našem případě a co vyjadřují? Zde je situace odlišná. Zaměříme se na vazby vznikající ohodnocením filmů. V tomto případě dochází k vazbě přes jiného zprostředkovatele. U fóra šlo o zprávu s nějakým obsahem. Obvykle nám při zkoumání centralit v této síti ani příliš nejde o to, co je obsahem těchto zpráv, ale maximálně to, na jaké téma se právě osoby baví. Podstatnější je počet zpráv a koho k sobě vážou. V případě naší sítě jsou zprostředkovateli vazeb filmy, které lidé hodnotí. Jakmile dva uživatelé ohodnotí jeden a ten samý film, dochází mezi nimi k vazbě. A my dokonce můžeme určit sílu této vazby na základě toho, jak se jejich názor shoduje. U písemné diskuze je toto velmi nečitelné, v našem případě naopak dokonale. Tím, že uživatelé ohodnotí film na nějaké stupnici, můžeme snadno najít nejen lidi v tomto případě se stejným vkusem, ale dokonce můžeme numericky vyjádřit i onu shodnost vkusu podle toho, jak moc se jejich hodnocení liší. Podobně jako jsou u fóra příspěvky na určité téma, zde můžeme filmy také třídit. A to podle mnoha kategorií. Například dle žánru, stáří, režiséra či země původu. A nyní se dostáváme k tomu nejpodstatnějšímu. Jaké informace

získáme zkoumáním centralit v těchto sítích? Počet příspěvků do fóra jednak vypovídá o jakési 'upovídanosti' uživatele, ale jak již bylo řečeno, můžeme také předpokládat, že častou diskusí na nějaké téma se také nějakým způsobem zvyšuje odbornost tohoto uživatele. Pokud se tedy na fórum podívame z pohledu centralit odvozených od degree centrality, jedná se o hledání jakéhosi nejdůležitějšího uživatele v síti. Nejaktivnějšího a teoreticky nejodbornějšího. Dokonce může záležet na tom, s kým uživatel komunikuje. Jestli s nějakým odborníkem či naopak z našeho pohledu nedůležitou osobou.

V naší síti ČSFD nedochází k výměně informací přímo. Zde je pojítkem mezi uživateli shodnost vkusu. Jakmile uživatel ohodnotí film, předává tímto tuto informaci ostatním. A naopak získává informace o vkusu jiných uživatelů. Co vlastně zjišťujeme skrze centrality týkající se vkusu? Ohodnocením jednoho filmu je uživatel okamžitě ve vztahu se všemi ostatními uživateli, kteří tento film také ohodnotili. Dá se logicky předpokládat, že čím více filmů uživatel ohodnotí, tím více vazeb bude mít a tedy se z našeho pohledu nějak zvýší jeho důležitost. To je pravda, ale jen do jisté míry. Ne každý film totiž přidává stejné množství vazeb. Snadno lze zjistit, že film který vidělo nejvíce českých diváků (předpokládame zde přímou úměru počtu diváků a počtu hodnocení) je Forrest Gump. Pokud uživatel ohodnotí takovýto mainstreamový film, vytvoří se obrovské množství vazeb. Takovýchto filmů je ale jen omezené množství. Čím více filmů uživatel ohodnotí, tím více mu docházejí takovéto mainstreamové filmy a musí tedy hodnotit málo známé filmy, takzvané artovky. A ty už mu mnoho vazeb nepřidají. V podstatě relativně zanedbatelné množství. Takže přímá úměra mezi počtem ohodnocených filmů a počtem vazeb se nedá předpokládat. Ačkoliv se tedy může zdát počet hodnocení některých uživatelů extrémní, v síti se to až tak výrazně neprojeví. V případě fóra jsme mohli u uživatele s nejvyšším ohodnocením předpokládat jeho nějakou důležitost či odbornost. Co ale můžeme předpokládat o uživateli s nejvyšším ohodnocením v našem případě? Co je to za člověka a je pro nás nějak zajímavý? Vazby mezi uživateli udávají shodnost vkusu. O uživateli s nejvyšším ohodnocením tedy můžeme předpokládat, že má nejshodnější vkus se zbytkem populace. Je to tedy typický filmový konzument. A to je určitě zajímavá informace. Především pro producenty z filmového průmyslu. Při úvahách jaký další film natočit se dívá do minulosti. Jaké starší filmy byly u diváků oblíbené? Na zjišťovaní tohoto je mnoho metrik a úhlů pohledu. Nejčastěji se na problém díváme ze dvou hledisek. Jednak jak vysoké průměrné hodnocení má film na takovýchto hodnotících databázích. Toto je velmi jednoduchá metrika, ale nemusí být zrovna nejrelevantnější. Druhé hledisko jsou samozřejmě peníze. Film nemusí mít žádnou převratnou uměleckou hodnotu, ale pokud na něj příjde hodně lidí a tedy hodně vydělá, je pro nás zajímavý. My se na problém díváme spíše z onoho hlediska kvality, ale trochu jinak. Tím, že najdeme uživatele s nejobecnějším vkusem, můžeme jeho vkus prozkoumat podle ohodnocených filmů a zjistit tak, co se průměrnému člověku líbí a co ne. Pokud zjistíme, že takovémuto uživateli se líbí třeba akční filmy s Brucem Willisem, není naškodu se zamyslet, jestli by nebylo výhodné natočit další obdobný film. Zatím jsme hovořili pouze o vazbách mezi uživateli tvořenými ohodnocením stejného filmu. Ale my máme k dispozici také jiný druh vazeb. Můžeme vytvořit síť, kde jsou vazby tvořeny shodností položek mezi oblíbenými. Kategorie výše popsaných oblíbených obsahují až deset titulů či tvůrců, které opět tvoří vazby s těmi

uživateli, kteří je mají také mezi oblíbenými. Rozdíl mezi vazbami je především takový, že u hodnocení určujeme váhu vazby na základě podobnosti hodnocení, kdežto u oblíbených jsou všechny vazby stejně silné a nás zajímá jen jejich počet. Ovšem i zde můžeme předpokládat, že jejich pořadí je ovlivněno vyšší/nižší oblíbeností a tedy se může jednat i o hodnocené vazby. Ale v tomto případě, kdy se jedná o pouhých deset nejvýznamnějších položek z velkého množství, je lepší je brát všechny jako stejně významné.

4.2 Praktická aplikace

Cílem zkoumání této sociální sítě bude porovnávání vkusu uživatelů ČSFD. Vazby mezi uživateli mohou být tvořeny na základě shodnosti ohodnocení nebo shodnosti v žebříčcích *oblíbených*. Máme tedy k dispozici dva způsoby jak sestrojit sítě.

4.2.1 Vytvoření vazeb

První způsob je takový, že jako spojovací články mezi uživateli použijeme jejich záznamy v *oblíbených*. Sestrojení těchto vazeb je poměrně jednoduché. Pouze projdeme konkrétní seznam *oblíbených* uživatele a najdeme všechny ostatní uživatele, kteří mají ve svých *oblíbených* tentýž záznam. Tím se vytvoří síť s vrcholy tvořenými uživateli a vazbami tvořenými společnými položkami v *oblíbených*. Ovšem síla vazby se také může různit. A to v případě, kdy mají dva uživatelé v dané kategorii *oblíbených* více než jeden společný záznam. Základní síla vazby se tedy pouze vynásobí počtem shodných záznamů. Může proto nabývat hodnot 1 až 10.

Druhým způsobem je využít hvězdičkového ohodnocení. Zde ovšem vyvstává zásadnější problém. Mezi dvěma uživateli vznikne velké množství různě silných vazeb, protože ohodnotili typicky mnoho společných titulů. Je tedy nutný složitější výpočet ohodnocení síly jedné konkrétní vazby mezi dvěma uživateli na základě zprůměrování shody jejich hodnocení titulů. Pokud se nalezne titul ohodnocený oběmi stranami, musí se vypočíst, jak moc se jejich hodnocení liší. Tím získáme sílu shody pro každou vazbu. Nakonec sílu těchto vazeb zprůměrujeme a tím jsme získali jednu finální vazbu mezi dvěma uživateli. Tyto výpočty budou ještě dále konkrétněji popsány.

4.2.2 Výpočetní náročnost

Vzhledem k tomu, že se operuje s obrovským množstvím dat, jsou výpočty shody ohodnocení výpočetně (a tedy časově) náročné. Pokud bychom chtěli sestrojit síť obsahující deset tisíc uživatelů, výpočet by trval neúnosně dlouhou dobu. Především u varianty s vazbami tvořenými na základě ohodnocení titulů. Jelikož má průměrný uživatel okolo 600 hodnocení a ke shodám dochází často, je potřeba zhruba 500 výpočtů pro porovnání dvou uživatelů. A když máme v síti deset tisíc uživatelů, je potřeba porovnat ohodnocení každého s každým, tedy 10000 krát 10000 porovnání děleno dvěma. Tedy přibližně $2,5 \times 10^{10}$ porovnání. Takovýto výpočet by trval velmi dlouho. Proto je nutné pracovat vždy s nějakou podmnožinou uživatelů, abychom získali menší sítě. Vytvoření takovýchto podmnožin je snadné. Například je možné vzít pouze uživatele z jednoho konkrétního

kraje nebo pouze uživatele s více než tisícem ohodnocení anebo podmnožinu uživatelů pseudonáhodně vybrat. Také je možnost nechat uživatele všechny, ale brát v potaz pouze nějaké konkrétní vazby splňující zadané podmínky. Třeba pouze akční filmy nebo pouze britské seriály atd. Tím získáme rozsahově mnohem rozumnější sítě.

5 Extrakce dat

Pro získání dat z portálu ČSFD byl vytvořen speciální program, který slouží ke stažení požadovaných dat z webu a jejich uložení do lokální databáze.

5.1 Účel aplikace

Aplikace slouží k automatizovanému procházení webových stránek portálu ČSFD a extrakci pro naše potřeby užitečných dat. Pro zkoumání této sociální sítě je potřeba získat dostatečný počet uživatelů a jejich údajů. Ne každý uživatel má vyplněny všechny možné informace. Pro účely zkoumání sítě jsou extrahovány tyto informace:

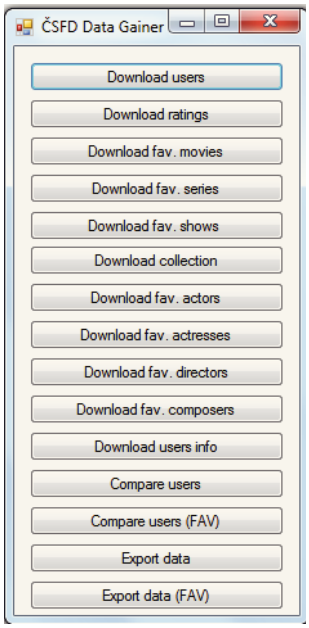
- Okres bydliště
- Hodnocení filmů
- Oblíbené filmy
- Oblíbené seriály
- Oblíbené pořady
- Oblíbení herci
- Oblíbené herečky
- Oblíbení režiséři
- Oblíbení skladatelé
- Filmy ve filmotéce

Tyto informace se uloží do lokální databáze pro další zpracování.

5.2 Návrh aplikace

Jádrem aplikace jsou dva objekty - jeden třídy *Parser* sloužící ke stahování a parsování webových stránek a druhý třídy *DB* pro operace s Microsoft SQL databází. Ihned po spuštění aplikace se vytvoří singleton třídy *DB* a ověří se spojení s lokální databází. V základním okně uživatelského rozhraní se zvolí požadovaná akce a poté se otevře modální okno s nastavením akce. Ve většině případů se jedná pouze o jednoduché zvolení intervalu uživatelů, jejichž data chceme stáhnout. Na obrázku 8 můžete vidět jednoduché uživatelské rozhraní (hlavní menu) programu.

Po spuštění stahování se vytvoří objekt třídy *Parser* a postupně se mu zadávají URL adresy na webové stránky, nad kterými má pracovat. Podle zvolené akce jsou volány příslušné metody objektu třídy *Parser*, které extrahují potřebná data ze zdrojových kódů webových stánek. Tato data se ukládají do kolekcí entit příslušného typu. Po načtení přichází na řadu práce s databází. Obvykle se jedná o insert (vložení) entit do databáze či jejich update (úprava). Celý tento princip bude ještě podrobněji popsán pro jednotlivé případy akcí.



Obrázek 8: Ukázka aplikace - hlavní menu

5.3 Entity

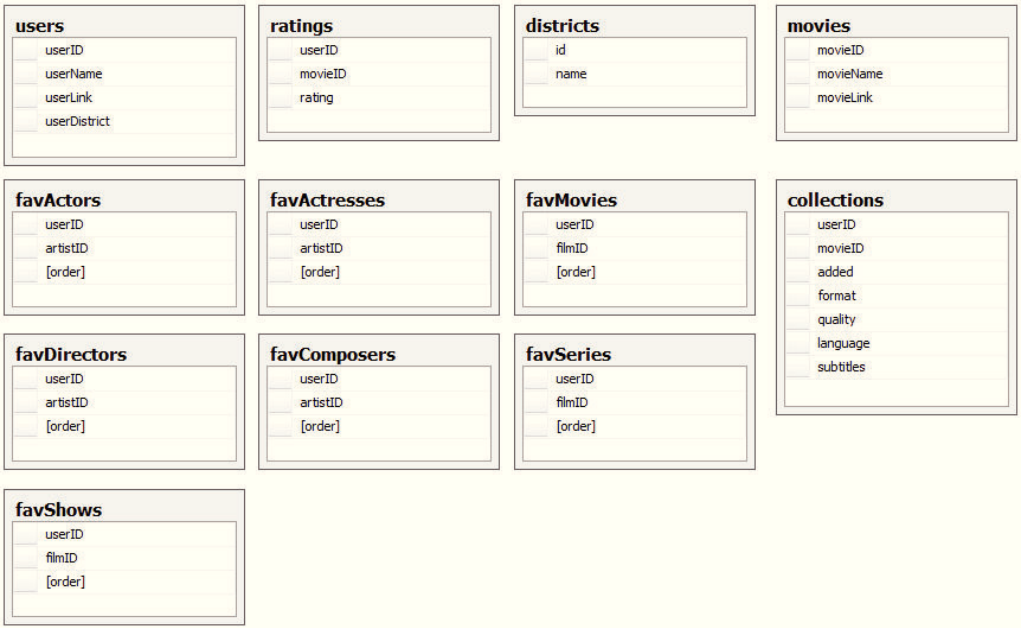
Pro reprezentaci dat v aplikaci slouží entity. Jsou to třídy pro různé typy objektů, se kterými se v programu pracuje. Tyto entity jsou následující:

- EArtist - tvůrce (režisér, herec atd.)
- ECollection - záznam o vlastnictví titulu ve sbírce
- EFavArtist - záznam přiřazující tvůrce do kategorie oblíbených konkrétního uživatele
- EFavFilm - záznam přiřazující titul do kategorie oblíbených konkrétního uživatele
- EMovie - film (ale také seriál či TV pořad)
- ERating - záznam přiřazující hodnocení titulu k uživateli
- EUser - uživatel

Tyto entity jsou jednoduché třídy obsahující atributy pro uložení potřebých informací (včetně jejich tzv. setterů a getterů) a konstruktor.

5.4 Databáze

K uložení dat se využívá Microsoft SQL databáze. Jednotlivé tabulky jsou vlastně předobrazy entit. Zde se již však budou entity tvůrců a titulů rozlišovat podle typu. Pro každý



Obrázek 9: Ukázka databáze – tabulky

typ (seriál, herec, skladatel atd.) je samostatná tabulka. Pro uchování dat je tedy vytvořeno celkem 12 tabulek. Na obrázku 9 jsou znázorněny všechny tabulky a jejich atributy.

5.5 Parser

Pro parsování webových stránek je vytvořena speciální třída *Parser*, která si zdrojový kód stánky stáhne z internetu a uloží do paměti. Pak jsou nad objektem typu *Parser* volány metody pro výtah požadovaných informací. Obsahuje metody jako např. `public List<EFavArtist> GetFavArtists(EUser user)`, která projde stránku se seznamem oblíbených tvůrců parametrem zadaného uživatele a pokud je má uvedeny, vrátí seznam jejich entit.

5.6 Práce s databází

Do databáze se přistupuje výhradně pomocí singletonu třídy DB, která obsahuje metody pro operace s jednotlivými tabulkami. Pro většinu tabulek jsou implementovány čtyři typy metod. Zde je příklad metod pro práci s uživateli.

- `public bool IsUserInDB(EUser user)` - zjistí zda je již daný uživatel v databázi
- `public void AddUser(EUser user)` - vloží uživatele do databáze
- `public void UpdateUser(EUser user)` - upraví uživatele v databázi podle údajů v předané entitě

- `public List<EUser> GetUsers(int from, int to)` - získá uživatele z databáze

Obdobné metody jsou implementovány i pro ostatní entity.

5.7 Akce stahování

V následujících kapitolách bude stručně popsána logika programu pro stahování vybraných, nejdůležitějších údajů z webu ČSFD.

5.7.1 Stažení uživatelů

Po zvolení stažení uživatelů se parseru zadá, aby stáhl stránky se seznamem uživatelů seřazených sestupně podle počtu hodnocení. Postupně se těmito stránkami prochází, uživatelé se ukládají do kolekce entit a poté vkládají do databáze.

```
Parser parser = new Parser(URLs.USERS_BY_RATING_START);

for (int page = 1; page <= 500; page++)
{
    if (page != 1) parser.LoadPage(URLs.USERS_BY_RATING_PAGE + page);
    foreach (EUser user in parser.GetUsers()) db.AddUser(user);
}
```

Výpis 1: Stažení uživatelů

5.7.2 Stažení hodnocení

Nejdříve se musí z databáze načíst uživatelé, pro které budeme hodnocení stahovat. U každého si zjistíme kolik stran s hodnocením jeho profil obsahuje a poté je dáme do kolekce entit a entity vložíme do databáze.

```
Parser parser = new Parser();

foreach (EUser user in db.GetUsers(From, To))
{
    string url = URLs.USER_START + user.UserURL + "hodnoceni/";
    parser.LoadPage(url);
    foreach (ERating rating in parser.GetRatings(user)) db.AddRating(rating);
}
```

Výpis 2: Stažení hodnocení

5.7.3 Stažení titulů a tvůrců

Zde se postupuje podobně jako u hodnocení, ale nyní je nutné rozlišovat o jaký typ záznamu se jedná, abychom načítali ze správné adresy a operovali se správnou tabulkou. Proměnná *Type* se nastaví při výběru akce.

```
Parser parser = new Parser();

foreach (EUser user in db.GetUsers(From, To))
{
    string url = URLs.USER_START + user.UserURL + URLEnding;
    parser.LoadPage(url);
    if (Type < Films.MOVIES)
    {
        foreach (EFavArtist favArtist in parser.GetFavArtists(user))
            db.AddFavArtist(favArtist, Type);
    }
    else
    {
        foreach (EFavFilm favFilm in parser.GetFavFilms(user))
            db.AddFavFilm(favFilm, Type);
    }
}
```

Výpis 3: Stažení titulů a tvůrců

6 Zpracování dat

Tato kapitola se zabývá získanými daty a jejich zpracováním tak, abychom na nich mohli vytvořit sítě a vypočítávat centrality. Data budou statisticky popsána, bude uveden postup jejich zpracování a jejich následného exportu do standardizovaného formátu.

6.1 Statistická analýza dat

Z portálu ČSFD byla extrahována data celkem deseti tisíců uživatelů. Jedná se o ty uživatele, kteří mají ve svém profilu nejvíce ohodnocených filmů a můžeme je tedy pokládat za nejaktivnější v síti. Databáze ČSFD eviduje k březnu roku 2013 přes 277 tisíc registrovaných uživatelů, drtivá většina z nich však nemá dostatečně vyplněné profily a jsou tedy z našeho pohledu nezajímaví.

6.1.1 Hodnocení filmů

Celkem bylo staženo 9658282 ohodnocení pro 132135 filmů. Uživatel s nejvíce ohodnoceními jich má 24421 a uživatel s nejméně ohodnoceními jich má 291. Průměrný uživatel v naší populaci má ohodnoceno 966 filmů. Medián je ovšem 526, což svědčí o tom, že několik málo uživatelů má abnormálně vysoký počet hodnocení. Na obrázku 10 je uveden histogram četnosti hodnocení. Z něj je zřejmé, že naprostá většina uživatelů má do tisíce ohodnocených filmů.

Další zajímavou statistikou je, jak lidé hodnotí. Na obrázku 11 jsou uvedeny četnosti velikosti jednolivých hodnocení - tedy kolik hvězdiček uživatelé filmům udělili. Lidé filmy ohodnocují nejčastěji čtyřmi hvězdičkami. Druhé nejčastější ohodnocení jsou tři hvězdičky. Naopak nejméně filmů v databázi má ohodnocení *odpad!*, tedy nula hvězdiček.

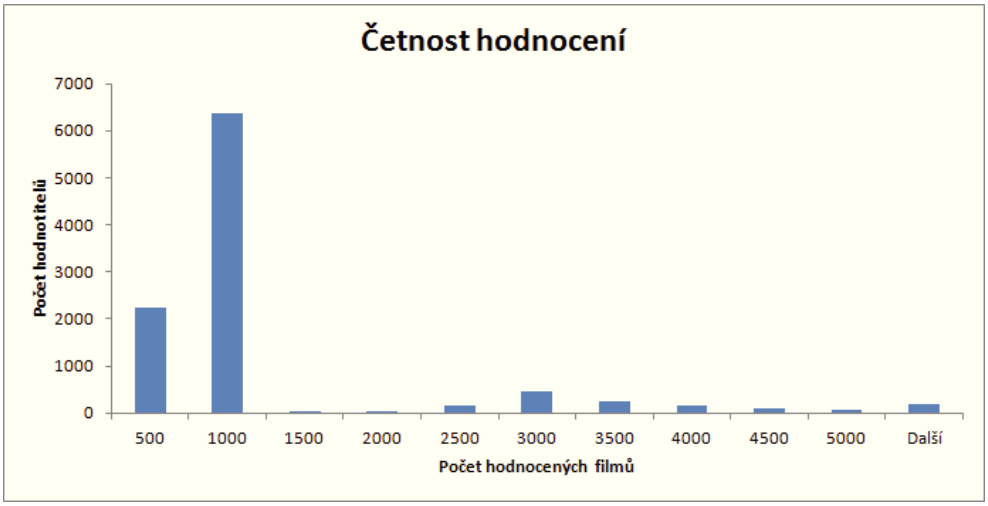
6.1.2 Kategorie *oblíbené*

Jak již bylo řečeno, ne každý uživatel má vyplněny kategorie oblíbených titulů a tvůrců. Z grafu na obrázku 12 je zřejmé, že lidé nejčastěji vyplňují své oblíbené filmy. Nejméně uživatelů má vyplněny své oblíbené skladatele. Avšak ne všichni mají v dané kategorii vyplněno všech deset míst. V tabulce 1 je uvedeno, jaký je průměr vyplněných pozic v TOP 10 jednotlivých kategorií. Když už někdo vyplní své oblíbené TV pořady, tak jich stejně neuvede mnoho. Nejpoctivěji lidé vyplňují kategorii oblíbených filmů.

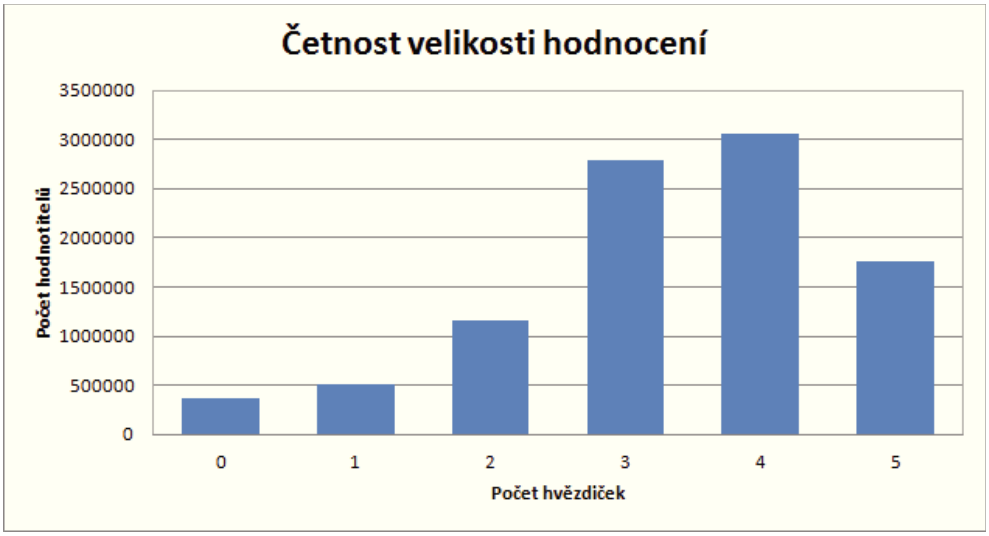
6.2 Výpočet shody hodnocení

Pro zjištění síly vazeb mezi uživateli je potřeba vypočíst koeficient shody jejich vkusu. Bude tedy vytvořena matice obsahující tento koeficient pro všechny dvojice uživatelů v síti. Spočíst takovouto matici pro více než tisíc uživatelů by bylo časově náročné, proto se budou matice vytvářet vždy pouze pro nějaké konkrétní podmnožiny uživatelů.

Nejdříve se načte seznam uživatelů, pro které budeme matici vypočítávat. Poté se sekvenčně berou uživatelé a porovnává se každý s každým. Pokud mezi nimi ještě není koeficient shody vypočten, je spuštěn výpočet. Při výpočtu koeficientu porovnáváme všechna



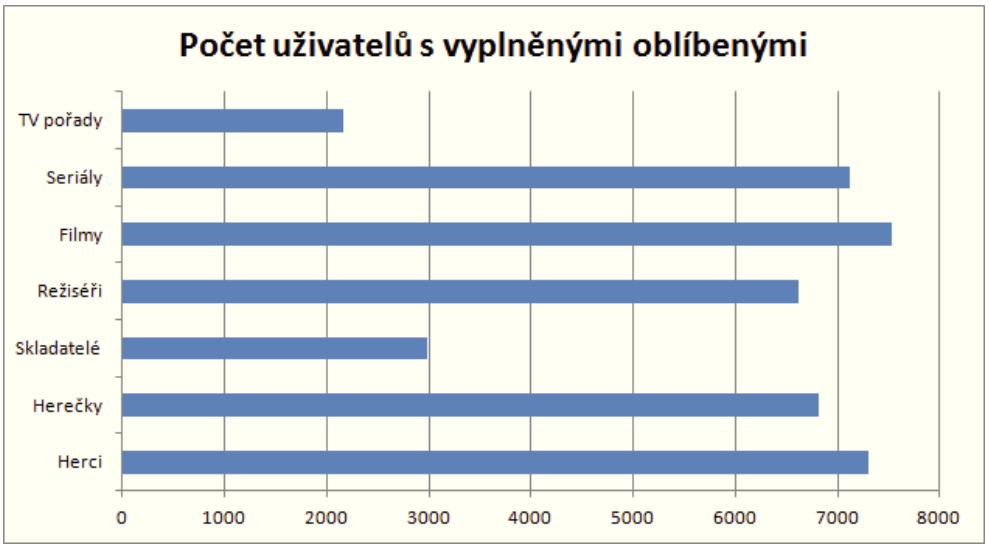
Obrázek 10: Histogram - četnosti hodnocení uživatelů



Obrázek 11: Četnosti velikosti hodnocení

Kategorie	Průměrné vyplnění
TV pořady	3,60
Seriály	7,31
Filmy	8,08
Režiséři	6,55
Skladatelé	4,93
Herečky	6,88
Herci	7,90

Tabulka 1: Tabulka průměrného vyplnění kategorií oblíbených



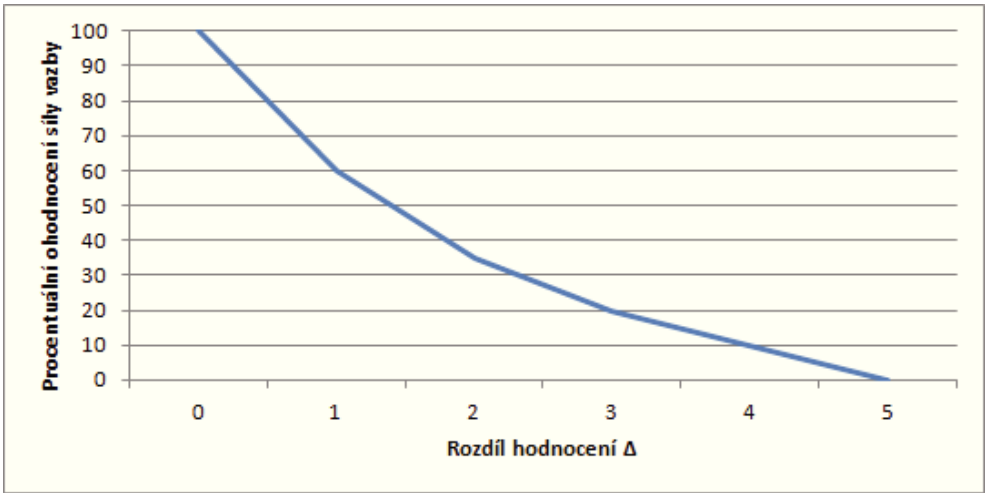
Obrázek 12: Počet uživatelů s vyplněnými *oblíbenými*

ohodnocení titulů jednoho uživatele s ohodnoceními druhého a dále pokračujeme pouze u těch titulů, které ohodnotili oba dva. U každé shody je potřeba určit rozdíl těchto dvou ohodnocení. Toto se provede jednoduchým odečtením počtu hvězdiček jednoho uživatele od počtu hvězdiček druhého a to v absolutní hodnotě podle vzorce 24.

$$\Delta = |H_1 - H_2| \tag{24}$$

Vzhledem k tomu, že počet hvězdiček nabývá hodnot 0 až 5, může také rozdíl být 0 pro shodné ohodnocení až 5 pro nejrozdílnější ohodnocení. Při výpočtu koeficientu shody budeme postupovat tak, že za každý takovýto rozdíl mezi dvěmi ohodnoceními připočteme ke koeficientu (začínajícímu na nule) určité číslo, vyjadřující sílu shody. Toto číslo by sice teoreticky mohlo být jednoduše onen rozdíl 0-5, ale v tomto případě by se na základě zkušenosti nereflektovala rozdílnost dobře. Vycházely by příliš si podobné výsledky a rozdíly mezi silami vazeb by tedy byly relativně malé. Pokud jeden uživatel ohodnotí film pěti hvězdičkami a druhý jen dvěmi, tak číselně je rozdíl jejich vkusu 60%, tedy 40% shoda. To je z lidského pohledu nerozumné. Rozdíl jejich vkusu je tak výrazný, že vyjádření jejich shody čtyřiceti procenty se zdá býti přehnané. Proto bylo rozložení transformováno, aby se rozdíly více projevily. Místo lineární funkce, byla použita přibližně exponenciální. V tomto případě nám o exaktní hodnoty příliš nejde, takže si může dovolit vhodné hodnoty sami určit podle vlastního expertního odhadu. Zvolíme tedy následující procentuální ohodnocení pro všechny Δ:

- Pro Δ = 0 bude ohodnocení $= 100$
- Pro Δ = 1 bude ohodnocení $= 60$
- Pro Δ = 2 bude ohodnocení $= 35$



Obrázek 13: Rozdělení ohodnocení shody vkusu

- Pro $\Delta = 3$ bude ohodnocení $= 20$
- Pro $\Delta = 4$ bude ohodnocení $= 10$
- Pro $\Delta = 5$, tedy naprostý rozdíl vkusu, bude ohodnocení nulové

Pokud tyto hodnoty vyneseme do grafu, získáme křivku na obrázku 13. Toto rozložení reflektuje reálný rozdíl vkusu lépe.

Vypočtené ohodnocení shody pro jednotlivé dvojice filmů se poté zprůměrují. Výsledkem bude koeficient v intervalu 0 až 100. Nula znamená, že tito dva uživatelé neshlédli žádné stejné filmy, nebo že se ve všech hodnoceních absolutně lišili. Naopak stovka vyjadřuje absolutní shodu ve všech filmech shlédnutými oběma. Je zde však zakomponováno ještě jedno omezení. Budeme brát v zřetel pouze vazby mezi uživateli, kteří zhlédli alespoň 50 stejných filmů. Jednak se nám tím sníží už tak velmi vysoký počet vazeb v grafu, ale především tím zamezíme zkreslujícím výsledkům. Vzorek méně než padesáti shod není příliš reprezentativní a může docházet k extrémním shodám jako 100%, přestože to s největší pravděpodobností neodpovídá realitě.

6.3 Export dat

Všechna získaná data jsou uložena v databázi, ale pro práci s nimi v jiných aplikacích je potřeba je exportovat do standardizovaného formátu. Byl zvolen formát GDF, který používá například The Graph Exploration System (GUESS) a Gephi [13, 14]. Tento formát má podobu coma separated file (CSV) souboru. Je rozdělen do dvou sekcí, jedné pro definici vrcholů grafu a druhé pro definici hran. Obě sekce začínají hlavičkovým řádkem, jež obsahuje definice sloupců. Níže je uveden jednoduchý příklad grafu ve formátu GDF. Tento graf má tři vrcholy a čtyři vážené hrany mezi nimi. Jednotlivé elementy (vrcholy i hrany) následují za definičním řádkem, každý element je na separátním řádku a sloupce

jsou odděleny čárkami. V tomto příkladu jsou vrcholy určeny pouze svým jménem a popiskem a hrany dvěmi vrcholy, které spojují a váhou oné hrany.

```
nodedef>name VARCHAR,label VARCHAR
s1,Site number 1
s2,Site number 2
s3,Site number 3
edgedef>node1 VARCHAR,node2 VARCHAR, weight DOUBLE
s1,s2,1.2341
s2,s3,0.453
s3,s2,2.34
s3,s1,0.871
```

Výpis 4: Příklad formátu GDF

7 Výpočty centralit

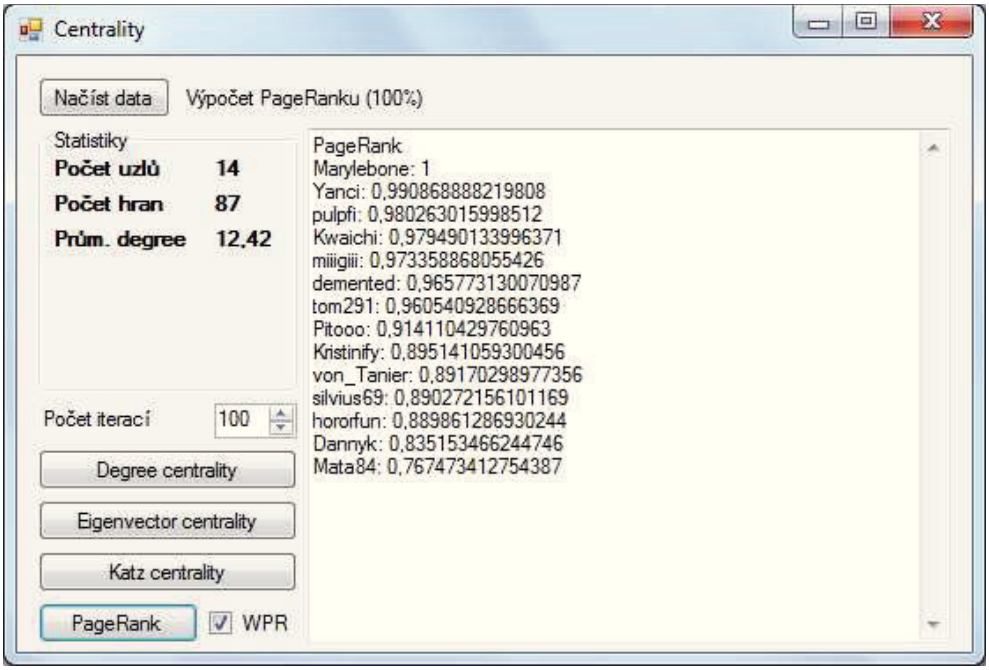
V této kapitole budou popsány procesy výpočtu jednotlivých zkoumaných centralit. Nejdříve bude představen program, sloužící k jejich kalkulaci, poté bude podrobně popsán algoritmus výpočtů těchto centralit. Výstupem programu nejsou absolutní čísla, jelikož jsou výsledné hodoty normalizovány do intervalu $\langle 0, 1 \rangle$, aby byly hodnoty centralit vzájemně snadno porovnatelné.

7.1 Program Centrality

Pro výpočty centralit odvozených od degree centrality při experimentování nad získanou sociální síti byl napsán vlastní program. Tento program je také napsán v programovacím jazyce C# a dokáže vypočítávat degree centrality, eigenvector centrality, Katz centrality a PageRank u grafů s váženými neorientovanými hranami. Uživatelské prostředí programu je na obrázku 14.

Program načítá data grafu z GDF souboru. Při čtení souboru spočte a zobrazí počet vrcholů a hran v grafu a všechny uloží do vnitřní paměti. Obdobně jako u předchozího programu, i zde ukádáme objekty do entit. V tomto případě máme dva druhy entit - `ENode` pro uložení vrcholu a `EEdge` pro uložení hrany. Po načtení grafu sítě můžeme započít výpočet požadované centrality. Pro všechny centrality kromě degree je relevantní nastavení maximálního počtu iterací. Ten určuje kolikrát maximálně se má provést upřesňující výpočet centrality. Povolené hodnoty jsou 1 až 1000. Čím více iterací nastavíme, tím přesnější výsledky získáme. Ovšem u větších sítí se může při nastavení vysoké hodnoty mírně prodloužit doba výpočtu.

Výpočty jednotlivých druhů centralit probíhají v separátních vláknech a výsledek je vždy vypsán do textového okna. Výpočet degree centralit je jednoduchý. Jednoduše se projde graf, vrchol za vrcholem a počítá se, kolik hran z/do něj vede. Výsledek pro každý uzel se uloží do atributu `Centrality` entity `ENode` a po skončení výpočtu se údaje vypíšou. U ostatních centralit je situace podstatně složitější. U eigenvector centrality a Katz centrality je potřeba nejdříve vypočítat vlastní čísla matice sousednosti. Graf sítě je zatím reprezentován seznamem vrcholů a hran v entitách. My však nyní graf převedeme do formátu matice sousednosti. Vznikne nám tedy matice `double[,] matrix`, kde jednotlivé elementy v matici představují váhu vazby mezi hranami. Vzhledem k povaze naší sociální sítě je tento výpočet poměrně časově náročný. Naše síť typicky obsahuje velmi velké množství hran mezi vrcholy. Každý vrchol v síti je spojen s většinou ostatních vrcholů. Protože musíme mnohokrát projít seznamem hran, jedná se o časově nejnáročnější část výpočtu centralit. Po vytvoření matice sousednosti je potřeba z ní vypočíst vlastní čísla. V tom nám vypomůže open source knihovna pro numerickou analýzu a zpracování dat ALGLIB [16]. Implementovaná metoda `rmatrixevd()`, které předáme naši matici sousednosti, vypočte vlastní čísla a my v daném vektoru vyhledáme nejvyšší vlastní číslo. Před začátkem samotného algoritmu výpočtu nastavíme centralitu všech vrcholů na 1. Poté spustíme výpočet pro určení jednotlivých druhů centralit, dle vzorců uvedených v kapitole 2. Ty budou dále podrobněji popsány. Stále přesnější centrality jsou ukládány do



Obrázek 14: Program Centrality

entit svých vrcholů a po ukončení výpočtu jsou normalizovány, seřazeny dle velikosti a vypsány do výstupního okna.

7.1.1 Výpočet eigenvector centrality

Před výpočtem nejdříve spočteme vlastní čísla z matice sousednosti a nalezneme nejvyšší z nich. Poté v zadaném počtu iterací provádíme následující výpočet. Pro každý vrchol vynásobíme centrality jeho sousedů (v první iteraci s přednastavenou centralitou rovnou jedné) s váhami vazeb uvedenými v matici sousednosti a všechny tyto součiny sečteme. Vzniklou sumu poté podělíme zjištěnou hodnotou nejvyššího vlastního čísla a tím získáme nový, přesnější odhad eigenvector centrality. Po proběhnutí požadovaného počtu iterací máme vypočtenou eigenvector centralitu pro každý vrchol v síti.

7.1.2 Výpočet Katz centrality

I v tomto případě je nutno určit vlastní čísla matice sousednosti. Tentokrát abychom mohli správně určit konstantu α. Dle [1] má být hodnota konstanty α blízká převrácené hodnotě nejvyššího vlastního čísla. Jak moc blízká, je na uvážení experimentátora. V našem případě byly zvoleny dva různé způsoby výpočtu α. Jeden podle vzorce 25 a druhý podle vzorce 26. První byl zvolen dle vlastního uvážení, druhý byl navrhnut z důvodu, že takto je konstanta α vypočítávána v programu NetworkX [17, 18]. Hodnota ev_{max} v následujících vzorcích je hodnota největšího nalezeného vlastního čísla.

$$\alpha = 0,9 \times \frac{1}{ev_{max}} \tag{25}$$

$$\alpha = \frac{1}{ev_{max} + 0,2} \tag{26}$$

Konstanta β byla nastavena β = 1. Postup výpočtu je dle vzorce 11. Tedy opět pro každý vrchol vynásobíme centrality jeho sousedů s váhami vazeb uvedenými v matici sousednosti a všechny tyto součiny sečteme. Nyní ale získanou sumu vynásobíme konstantou α a přičteme konstantu β. Jelikož je $\alpha < ev_{max}$, měl by algoritmus konvergovat [1]. Tento výpočet se opět opakuje tolikrát, na kolik je nastavena maximální iterace.

7.1.3 Výpočet PageRanku

Výpočet PageRanku je mírně odlišný. Nyní již nepotřebujeme vypočítávat hodnoty vlastních čísel matice sousednosti. Výpočet probíhá podle vzorce 13. Je obdobný jako v předchozím případě, jen nyní při výpočtu sumy centralit sousedů vrcholu podělíme každou centralitu daného souseda stupněm jeho degree centrality. Zde si musíme pouze pohlídat, abychom nedělili nulou. Pokud má vrchol stupeň degree centrality roven nule, změníme jej na 1. Na výpočet to nebude mít žádný vliv [1]. Ovšem i zde je nutné náležitě nastavit konstanty α a β. Google i NetworkX nastavuje konstantu α = 0,85 [1, 18], učiníme tak tedy i my. Konstantu β nastavíme podle vzorce 15, což je nejstandardnější způsob. Program Centrality dokáže vypočítávat dva druhy centralit. Při nezatržení checkboxu *WPR* se vypočítává nevážený PageRank podle vzorce 13 a při jeho zatržení se vypočítává vážený PageRank podle vzorce 20.

7.2 Vytvoření sítí pro experimenty

Pro účely experimentů byly vytořeny různé podsítě sociální sítě uživatelů ČSFD. Populace byla vždy omezena, jelikož počítání se všemi uživateli portálu ČSFD by bylo časově velmi náročné a výsledné sítě by nešly rozumně graficky znázornit. Uživatelé byli vybíráni na základě jejich bydliště a vazby byly tvořeny dle odlišných, níže uvedených kritérií. Byl brán ohled na zajímavost sítě jak z pohledu praktického, abychom získali zajímavé informace, tak pohledu akademického, abychom mohli porovnat jednotlivé centrality. Vytvořené sítě jsou následující:

- Uživatelé z města Ružomberok - malá síť čtrnácti uživatelů
  - Vazby vytvořené na základě hvězdičkového hodnocení filmů
- Uživatelé z města Ostravy - velká síť dvěstětřiceti uživatelů
  - Vazby vytvořené na základě hvězdičkového hodnocení filmů
  - Vazby vytvořené na základě filmů v *oblíbených*
  - Vazby vytvořené na základě akčních filmů v *oblíbených*

- Vazby vytvořené na základě seriálů v *oblíbených*
- Vazby vytvořené na základě skladatelů v *oblíbených*

Na síti uživatelů z města Ružomberok se dobře projeví různé vlastnosti zkoumaných centralit a také grafické znázornění této malé sítě bude přehledné. Ostrava byla zvolena z důvodu, že je pro nás zajímavá z hlediska lokace, ale také většímu množství uživatelů. Grafické znázornění takto velké sítě nebude tak vypovídající jako v předchozím případě, ale zase získáme pro nás zajímavější informace o vkusu uživatelů z našeho okolí. Místo toho, abychom zkoumali větší množství naprosto odlišných sítí, zaměříme se na uživatele z Ostravy a prozkoumáme je z různých hledisek. Počítání centralit na těchto podsítích bude přiměřeně rychlé a přitom stále smysluplné. Degree centralita je v obou případech vypočítána velmi rychle, v řádu desetin vteřiny. Výpočet eigenvector centrality, Katz centrality i PageRanku trvá přibližně stejně, a to řádově desetiny vteřiny pro síť uživatelů z města Ružomberok a jednotky vteřin pro uživatele z města Ostravy, podle počtu vazeb v dané síti. Z důvodu vysokého provazbení sítě používáme místo typičtější řídké matice sousednosti plnou matici sousednosti a právě konstrukce této matice je z celého výpočtu časově nejnáročnější částí, jak již bylo zmíněno v kapitole 7.1.

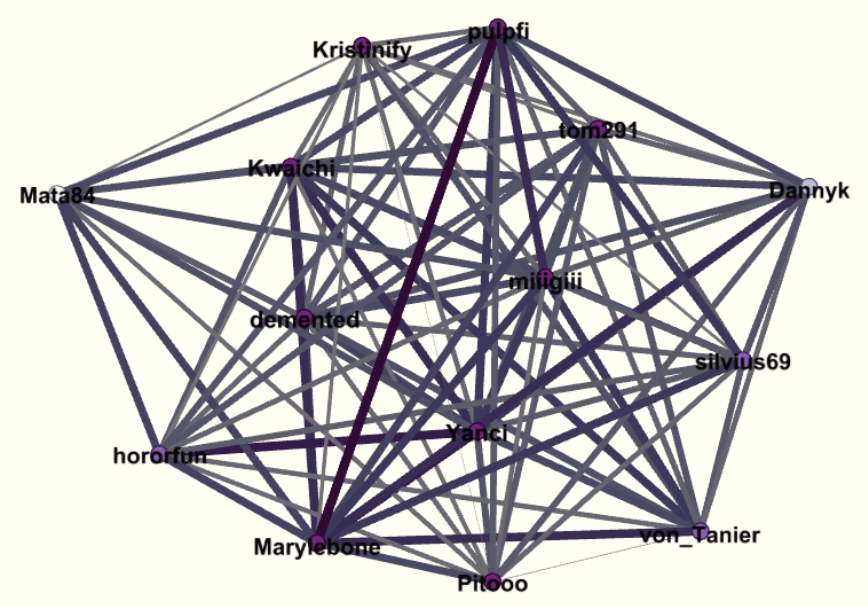
7.3 Analýza shodnosti ohodnocení

V této kapitole budou rozebrány výsledky různých experimentů provedených vytvořeným programem Centrality na rozličných podsítích sociální sítě uživatelů ČSFD. Pokud nebude uvedeno jinak, budou konstanty nastaveny následovně. U Katz centrality je α nastavena dle vzorce 25 a $\beta = 1$, u PageRanku $\alpha = 0,85$ a β dle vzorce 15. Správnost výsledných hodnot degree centrality a PageRanku byla zkontrolována podle programu Gephi a hodnoty eigenvector centrality podle programu NetworkX.

7.3.1 Síť uživatelů z města Ružomberok

Nejdříve byly vypočteny centrality uživatelů ze slovenského města Ružomberok. Výsledky jsou vypsány v tabulce 2. Hodnoty v této tabulce jsou seřazeny sestupně podle eigenvector centrality. Tato síť je tvořena celkem čtrnácti uživateli (vrcholy), mezi nimiž je 87 ohodnocených vztahů (hran). Síť je graficky znázorněna na obrázku 15, vygenerovaném programem Gephi. Je zde čtrnáct kruhů, znázorňujících vrcholy a 87 čar, znázorňujících hrany. Šířka a barva hrany vyjadřuje sílu (váhu) vazby. Čím hrubší hrana, tím silnější vazba, čím fialovější hrana, tím silnější vazba. Barvou je také znázorněna degree centralita, opět čím fialovější vrchol je, tím větší má centralitu. Tato pravidla platí i pro všechna další grafická znázornění sítí v práci.

Na tomto malém vzorku je dobře vidět, že degree centralita je kvůli své jednoduchosti poměrně nepřesnou centralitou. Ze čtrnácti vrcholů má devět stejný počet hran (konkrétně 13) a tedy všech devět je ohodnoceno stejnou degree centralitou. Třináct hran na vrchol je v této síti maximum a není tedy zřejmé, kterého z těchto devíti uživatelů můžeme považovat za nejdůležitějšího. Nejmenší degree centralitu má uživatel *Mata84*. Získal pouhých 0,7692 oproti maximální centralitě 1. To je samozřejmě způsobeno tím, že jeho reprezentující

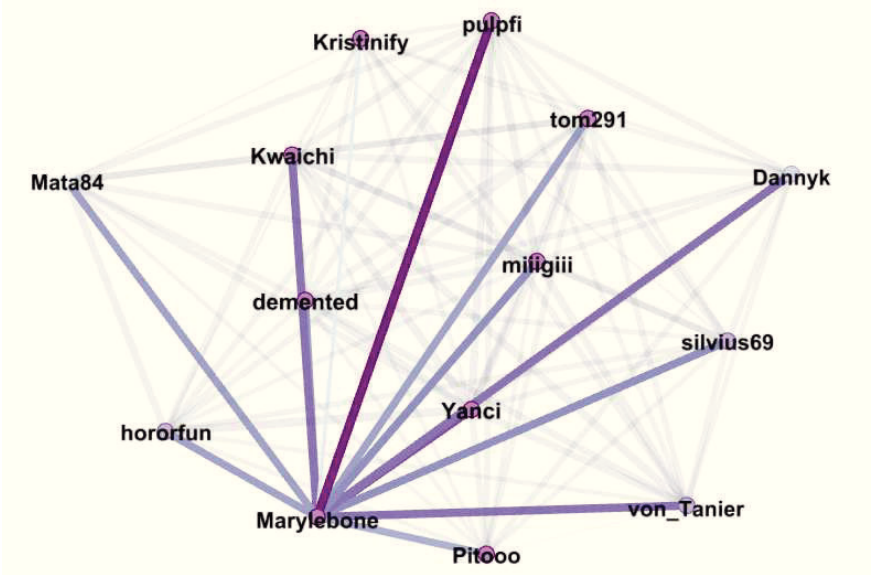


Obrázek 15: Graf sítě uživatelů z města Ružomberok

vrchol má pouze 10 hran. V dalším sloupci jsou hodnoty eigenvector centrality. Zde jsou již výsledky mnohem různorodější. Nejvyšší eigenvector centralitu má uživatel *Marylebone*. Znamená to, že je v síti z tohoto pohledu nejdůležitější. Díky degree centralitě víme, že tento vrchol má stejně hran jako osm dalších. Projevila se zde tedy ona vlastnost eigenvecotor centrality, že nezáleží pouze na počtu hran, ale také na tom, jak důležité vrcholy jsou v sousedství tohoto vrcholu. Na obrázku 16 je onen vrchol a jeho hrany zvýrazněny. Vidíme, že je spojen se všemi ostatními vrcholy grafu a všechny vazby jsou silné, jelikož čáry znázorňující hrany jsou hrubé. Na obrázku 16 vede nejtenčí hrana do vrcholu *Kristinify*. Tato síť je specifická tím, že všechny vrcholy mají téměř stejné sousedy. Proto se zde výše zmíněná vlastnost eigenvector centrality neprojevuje příliš silně. V tabulce 2 vidíme, že hodnota eigenvector centrality se u těch vrcholů, které mají shodnou degree centralitu, příliš neliší.

Ještě se zaměříme na uživatele *Mata84*, který má nejmenší i eigenvector centralitu. Na obrázku 17 jsou znázorněny jeho vazby. Vidíme, že mu opět škodí především to, že je spojen pouze s deseti jinými vrcholy. Toto ho tak silně devaluje, že ani nepomáhá to, že vazby jsou poměrně silné (čáry hran jsou relativně hrubé). S těmi uživateli, s kterými má vazbu, se ve filmech názorově poměrně shodne. Ale zřejmě sleduje příliš netradiční filmy vzhledem k dané skupině. To, že nemá vazbu s třemi uživateli, je totiž způsobeno tím, že v jejich ohodnoceních nebyl dostatečný počet styčných filmů na to, aby se vytvořila vypovídající vazba.

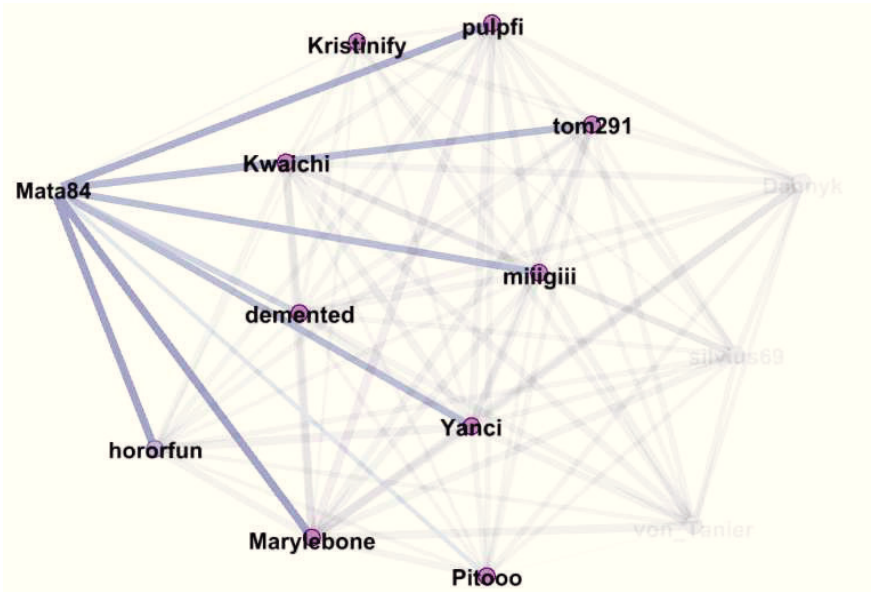
Pokud v tabulce 2 porovnáme hodnoty eigenvector centrality a Katz centrality, vidíme, že se příliš neliší. Z toho můžeme usuzovat, že přinejmenším pro síť našeho typu jsou tyto metriky velmi podobné. Pořadí uživatelů dle velikosti Katz centrality je shodné jako pořadí



Obrázek 16: Graf sítě uživatelů z města Ružomberok - zvýraznění vrcholu *Marylebone*

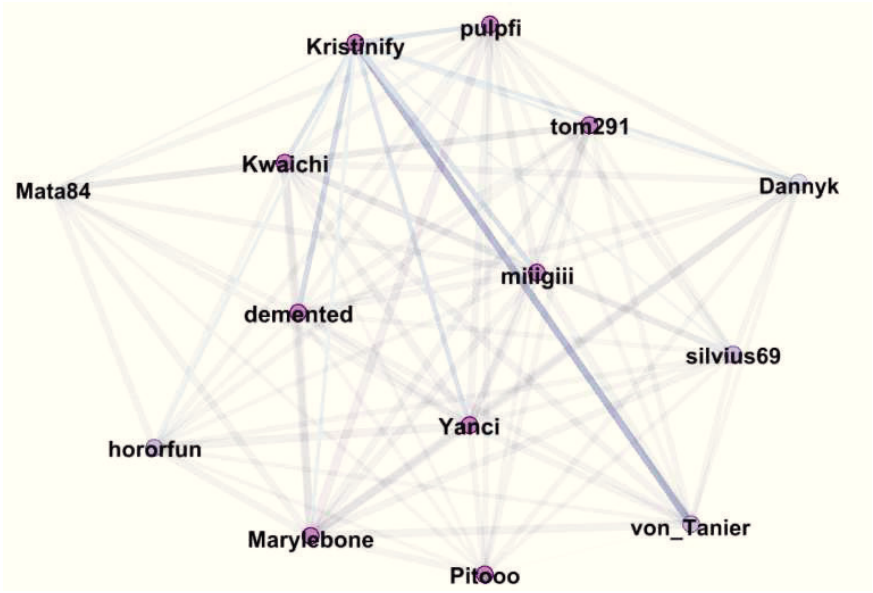
Uživatel	Degree	Eigenvector	Katz	PageRank	Nevážený PR
Marylebone	1	1	1	1	1
Yanci	1	0,9905	0,9913	0,9909	1
pulpfi	1	0,9791	0,9809	0,9803	1
Kwaichi	1	0,9783	0,9802	0,9795	1
miiigiii	1	0,9723	0,9747	0,9733	1
demented	1	0,9641	0,9672	0,9658	1
tom291	1	0,9579	0,9615	0,9605	1
Pitooo	1	0,9092	0,917	0,9141	1
von_Tanier	0,9231	0,8992	0,9069	0,8917	0,9315
silvius69	0,9231	0,8969	0,9049	0,8903	0,9315
hororfun	0,9231	0,8891	0,8983	0,8899	0,9325
Kristinify	1	0,8885	0,8981	0,8951	1
Dannyk	0,8462	0,8412	0,8538	0,8352	0,8647
Mata84	0,7692	0,7634	0,7833	0,7675	0,7987

Tabulka 2: Tabulka centralit uživatelů z města Ružomberok



Obrázek 17: Graf sítě uživatelů z města Ružomberok - zvýraznění vrcholu *Mata84*

dle velikosti eigenvector centrality. Pro tento první experiment byl vypočten nejen vážený PageRank, ale také pro srovnání i nevážený. Vidíme, že hodnoty neváženého PageRanku jsou velmi podobné degree centralitě. Odchylku vidíme u uživatele *hororfun*. Má spolu s uživateli *silvius69* a *von_Trainer* shodnou degree centralitu, ale jeho PageRank se liší. Jeho mírně vyšší hodnota PageRanku je způsobena jedním odlišným sousedem. Všechny tři tyto vyrcholy reprezentující uživatele mají dvanáct hran vedoucích k sousedům. *silvius69* a *von_Trainer* mají stejné sousedy a jeden z nich je *Dannyk*. *hororfun* ovšem tohoto souseda nemá a místo něho má vazbu s uživatelem *Mata84*. V tabulce 2 vidíme, že *Mata84* má nižší degree centralitu než *Dannyk* a proto *Mata84* přináší do vrcholu *hororfun* větší PageRank. U váženého PageRanku vyšly různorodější hodnoty. Již záleží na vahách hran, které se velmi různí, takže žádné dvě hodnoty váženého PageRanku nejsou shodné. V několika případech se dokonce změnilo pořadí. Například uživatel *Kristinify* není na dvanáctém místě jako v případě eigenvector centrality, ale na devátém. Má nejmenší vážený PageRank z vrcholů s nejvyšší degree centralitou. Pokud se na vrchol reprezentující uživatele *Kristinify* podíváme na obrázku 18 vidíme, že má několik velmi slabých vazeb. Jeho vazby jsou nejslabší z vrcholu s degree centralitou rovnou jedné a proto je mezi nimi poslední. Tyto slabé vazby se u eigenvector centrality a Katz centrality projevily výrazněji než u PageRanku. PageRank tedy přikládá větší váhu počtu hran než ohodnocení hran narozdíl od eigenvector centrality a Katz centrality. Vidíme, že vážený PageRank nám v našich sítích dává přesnější výsledky, proto už v následujících experimentech nebudeme nevážený PageRank počítat. Všechny následující sítě mají také ohodnocené hrany, takže výsledky hodnoceného PageRanku budou více odpovídající realitě.



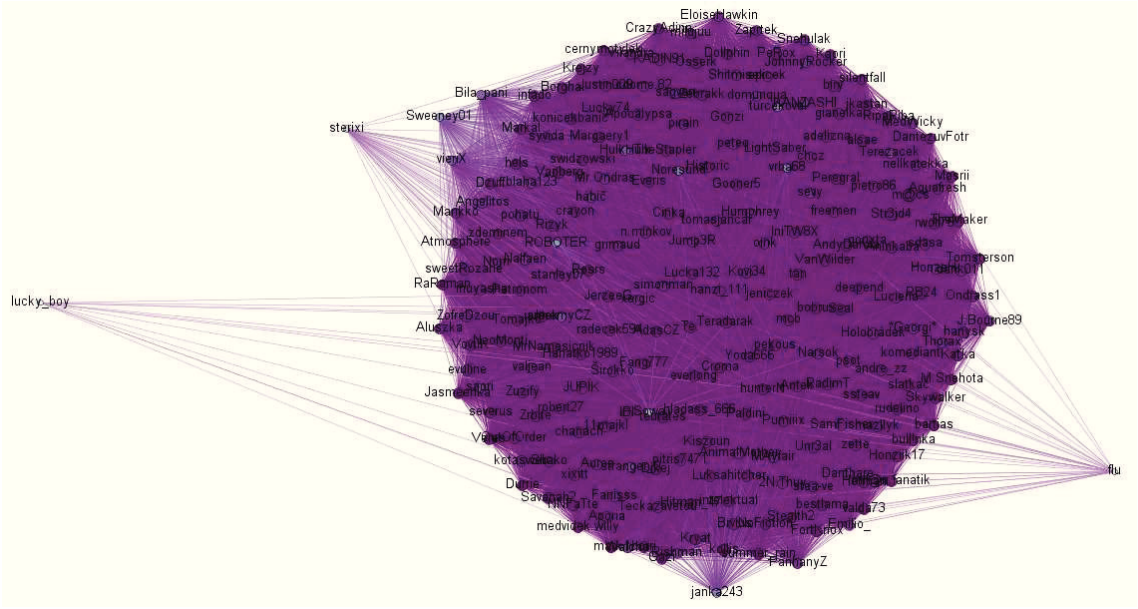
Obrázek 18: Graf sítě uživatelů z města Ružomberok - zvýraznění vrcholu *Kristinify*

7.3.2 Síť uživatelů z města Ostravy

Druhou zkoumanou sítí je síť uživatelů z Ostravy. Tato síť obsahuje 230 uživatelů (vrcholů) a 24489 hodnocených vztahů (hran). Na síť opět aplikujeme algoritmy pro výpočet centralit a rozebereme výsledky. Vzhledem k vysokému počtu hran není grafické znázornění v tomto případě ideální, jelikož graf je značně nepřehledný. Na obrázku 19 je tento graf pro ilustraci vykreslen. V tabulce 3 jsou uvedeny vypočtené hodnoty centralit pro deset uživatelů s největší eigenvector centralitou a deset uživatelů s nejmenší eigenvector centralitou.

Ve všech čtyřech případech získal nejvyšší centralitu uživatel *Brvius*. Tento uživatel má nejtypičtější vkus ze všech v této síti. V tom mu pomáhá skutečnost, že ohodnotil velký počet filmů (3756). Má vazby se všemi ostatními uživateli a tedy samozřejmě nejvyšší možnou degree centralitu. Celkem je v síti dvanáct vrcholů s maximální degree centralitou. *Brvius* má z nich nejvyšší eigenvector centralitu, Katz centralitu i PageRank, jelikož jeho vazby se sousedy jsou nejsilnější. V tabulce 4 jsou uvedeny hodnoty průměrného ohodnocení hran vrcholů a jejich pořadí dle eigenvector centrality u vrcholů s nejvyšší (shodnou) degree centralitou. Lze vidět, že v případě, kdy vrcholy mají jako sousedy všechny ostatní vrcholy v síti, záleží při výpočtu eigenvector centrality především na velikosti ohodnocení hran vedoucích k těmto sousedům. Tedy samotná vysoká degree centralita vrcholu nezaručuje vysokou eigenvector centralitu, ale obdobný vliv má také ohodnocení hran.

Při pohledu na spodek tabulky 3 vidíme velmi malé centrality u uživatele *lucky_boy*. Jeho výjimečnost je zřejmá už z obrázku 19. Tento uživatel má malou především degree centralitu. Jej reprezentující vrchol má pouze 13 hran vedoucích k jiným vrcholům. Hodnoty těchto vah jsou poměrně silné (s průměrem 91,39), což způsobuje, že ostatní



Obrázek 19: Graf sítě uživatelů z města Ostravy - hrany dle hodnocení

centrality jsou tedy výrazně vyšší než degree centralita. Eigenvector centralita je vyšší jen o 8%, ale Katz centralita už o 163% a PageRank dokonce o 397%. Co způsobilo, že uživatel *lucky_boy* má tak malou degree centralitu? Pokud se na profil tohoto uživatele podíváme podrobněji, zjistíme, že sám o sobě prohlašuje, že sleduje takřka výhradně klasické Hollywoodské filmy staršího data vzniku. A vskutku, jeho vkus je extrémně neobvyklý. Většinu jím ohodnocených filmů vidělo pouze minimum jiných uživatelů a vytvořilo se tedy pouze 13 vazeb s ostatími uživateli. Pro srovnání si ještě přiblížíme vrchol reprezentující uživatele *turcekoval*. Tento vrchol má 157 hran s průměrem vah 74,62. Počet hran je výrazně vyšší a je tedy podle očekávání také vyšší degree centralita. Váhy hran jsou však v tomto případě znatelně nižší a to se projevilo u ostatních centralit. Jsou dokonce nižší než degree centralita. Eigenvector centralita je nižší o 14%, Katz centralita o 9% a PageRank také o přibližně 9%. Je tedy zřejmé, že hodnoty vah hran výrazně ovlivňují výsledné hodnoty centralit. Je ovšem obtížné říci, která z těchto čtyř druhů centralit je v našem případě nejvíce vypovídající.

7.4 Analýza shodnosti oblíbených kategorií

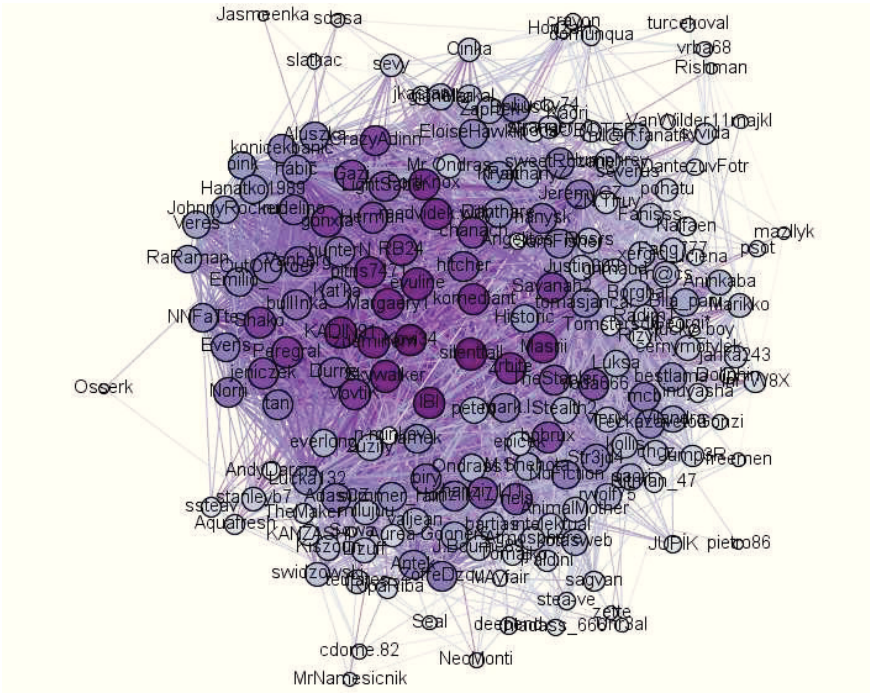
U *oblíbených* je problém v tom, že ne všichni uživatelé je mají vyplněné. Budeme tedy zkoumat větší síť, abychom měli dostatek údajů. Pro srovnání vypočteme centrality opět v síti uživatelů z Ostravy. Graf tedy bude mít stejné vrcholy, ale nyní jiné hrany. Z kategorií oblíbených se nejdříve zaměříme na oblíbené filmy. Váha hrany se bude tentokrát pohybovat v rozmezí 1 až 10, podle toho, kolik shodných filmů uživatelé mají v *oblíbených*.

Uživatel	Degree	Eigenvector	Katz	PageRank
Brvius	1	1	1	1
Durrie	0,9738	0,9995	0,9983	0,9757
AdasCZ	0,9869	0,9991	0,9982	0,9773
Historic	0,9869	0,9964	0,9958	0,9751
Gazi	1	0,9963	0,9967	0,9964
Tomajko	0,9956	0,9956	0,9956	0,9836
jeniczek	0,9825	0,993	0,9925	0,9705
mark.l	0,9782	0,9925	0,9918	0,9478
stanleyb7	0,9651	0,9915	0,9906	0,9621
Kryat	1	0,9911	0,9919	0,9964
Noresund	0,6507	0,6364	0,6688	0,6583
vrba68	0,6419	0,6133	0,6481	0,6417
turcekoval	0,6856	0,5866	0,6242	0,6218
janka243	0,6157	0,567	0,6061	0,6011
ROBOTER	0,5022	0,5175	0,5609	0,5557
Sweeney01	0,5109	0,5024	0,5478	0,5476
Sowa	0,3974	0,4078	0,4618	0,4638
sterixi	0,2271	0,2184	0,2918	0,3125
flu	0,166	0,1694	0,247	0,2724
lucky_boy	0,0568	0,0613	0,1491	0,1849

Tabulka 3: Tabulka zkoumaných centralit uživatelů z města Ostravy

Uživatel	Průměrné ohodnocení hrany	Pořadí dle eigenvector centrality
Brvius	89,36	1.
Gazi	89,02	5.
Kryat	88,55	10.
everlong	88,14	19.
mcb	87,94	28.
hunterN	87,07	58.
deepend	86,54	75.
iamek	86,10	82.
Luksa	86,13	83.
kollis	85,80	95.
Seal	84,83	121.
Croma	81,20	179.

Tabulka 4: Uživatelé z Ostravy s nejvyšší degree centralitou



Obrázek 20: Graf sítě uživatelů z města Ostravy - hrany dle oblíbených filmů

7.4.1 Oblíbené filmy

Graf sítě s vazbami na základě shodnosti v oblíbených filmech se skládá z 230 vrcholů a 3712 hran. Grafické znázornění sítě je na obrázku 20. Vypočtené centrality pro deset uživatelů s nejvyšší eigenvector centralitou a deset uživatelů s nejmenší eigenvector centralitou jsou v tabulce 5. Údaje jsou opět seřazeny podle velikosti eigenvector centrality. Tentokrát jsou vynecháni uživatelé, kteří nemají oblíbené filmy vyplněné a mají tedy nulovou degree centralitu.

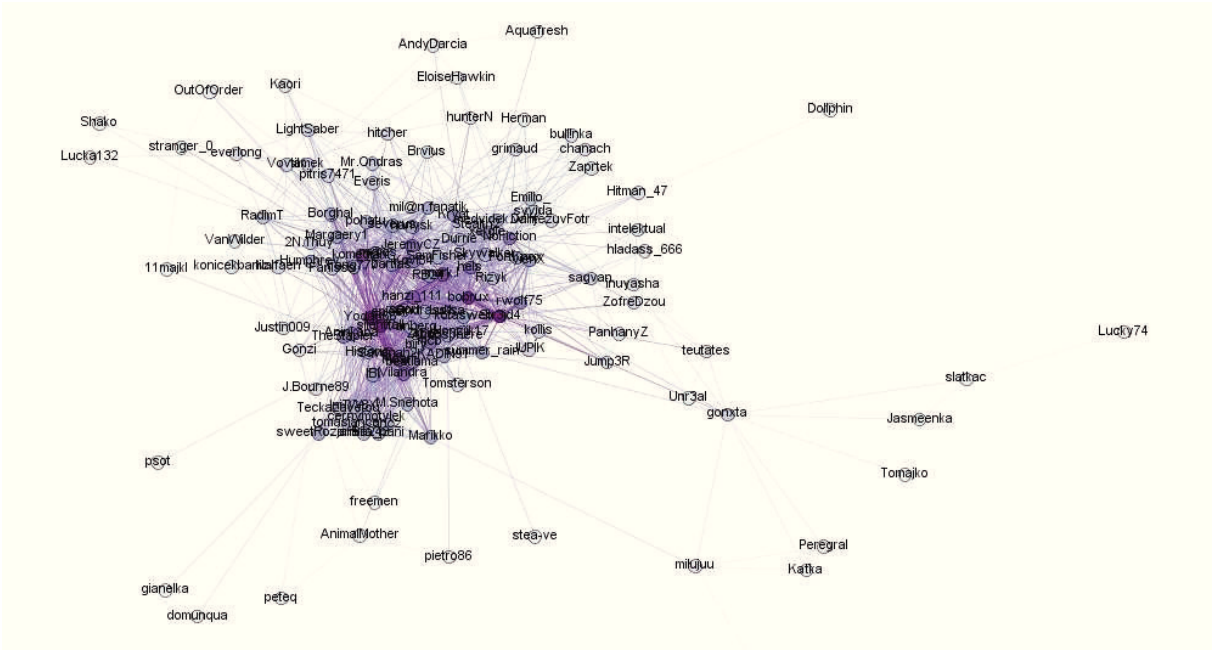
Zde vyšly hodnoty odlišně, ale několik uživatelů se ve výběru opakuje. Například uživatel *lucky_boy* se opět nachází na konci tabulky. Jeho odlišnost ve vkusu se tedy projevila i zde. Zajímavě také vyšly hodnoty PageRanku. I uživatelé s téměř nulovou degree centralitou získali poměrně vysoký PageRank v porovnání s eigenvector či Katz centralitou. To je způsobeno tím, že PageRank poměrně významně hodnotí i vrcholy s velmi malou degree centralitou, jelikož počítá s jakýmsi náhodným odskokem do těchto vrcholů.

7.4.2 Oblíbené akční filmy

Pro porovnání byla vytvořena další síť, kde vazby mezi uživateli tvoří společné oblíbené filmy, ale tentokrát byla omezena pouze na akční filmy. Sestrojení sítě bylo provedeno stejným způsobem, ale pokud shodný film nebyl akčního žánru, byl ignorován. Tím se značně snížil počet vazeb mezi uživateli na méně než třetinu. Tato nová síť má tedy 1141 hran pro

Uživatel	Degree	Eigenvector	Katz	PageRank
zdeminem	0,9072	1	1	1
Kovi34	1	0,8866	0,8991	0,9687
silentfall	0,9794	0,8569	0,8736	0,9505
medvidek.willy	0,8557	0,8197	0,8315	0,8745
KADIN91	0,9175	0,8169	0,8288	0,8795
Masrii	0,8866	0,801	0,8212	0,9109
FortKnox	0,8557	0,7721	0,7859	0,8312
RB24	0,8041	0,7409	0,7519	0,7508
Skywalker	0,9072	0,7368	0,7539	0,7973
Zrbite	0,8557	0,7269	0,7526	0,8561
mazllyk	0,0412	0,0016	0,0646	0,1131
Lucky74	0,0412	0,0113	0,0641	0,1132
Lucky_boy	0,0309	0,011	0,0632	0,0876
Rishman	0,0309	0,0106	0,063	0,0969
Osserk	0,0206	0,0096	0,0614	0,0757
turcekoval	0,0515	0,0092	0,0631	0,1139
Seal	0,0619	0,0091	0,0639	0,1158
Jasmeenka	0,0206	0,0085	0,0611	0,1095
Sowa	0,0206	0,008	0,0607	0,0841
pietro86	0,0206	0,0047	0,0574	0,0770

Tabulka 5: Tabulka centralit uživatelů z města Ostravy - oblíbené filmy



Obrázek 21: Graf sítě uživatelů z města Ostravy - hrany dle oblíbených akčních filmů

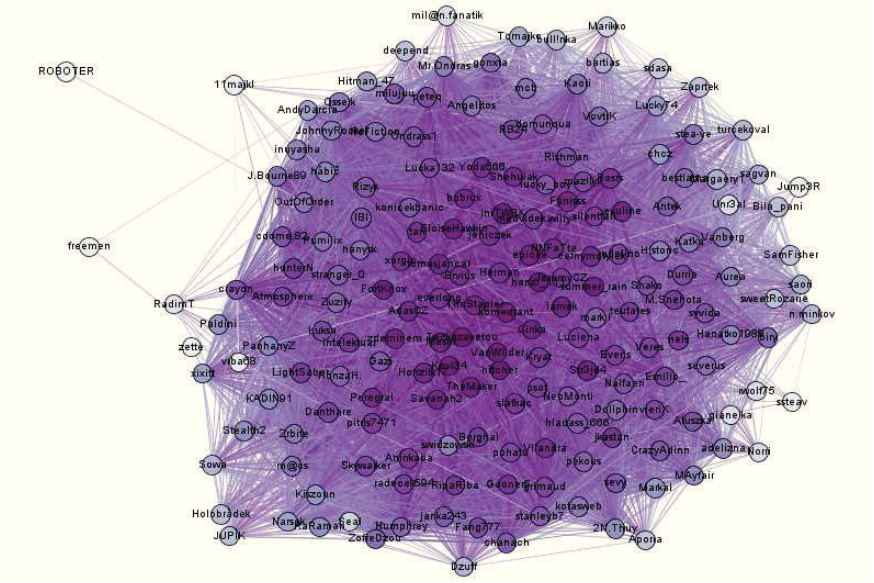
230 uživatelů a je graficky znázorněna na obrázku 21. V tabulce 6 jsou uvedeny vypočtené centrality pro uživatele vybrané podle stejného kritéria jako v předchozím případě. Graf této sítě už není tak kompaktní, jako při počítání bez omezení na žánr. Vyskytlo se zde více uživatelů odchylujících se od průměru typického diváka. V tabulce podle všech měřených centrality dominuje uživatel *Yoda666*, který jednoznačně reprezentuje nejtypičtějšího konzumenta akčních filmů v populaci. Většina jeho zvolených oblíbených filmů jsou slavné akční hity jako Hvězdné války, Pán prstenů či filmy s Batmanem. Při pohledu na spodní část tabulky vidíme naopak netypické diváky z pohledu akčních filmů. Například uživatelka *Jasmeenka* má mezi svými oblíbenými filmy víceméně jen nepříliš známé horory jako Dům tisíce mrtvol či Mučedníci, jejichž názvy nejspíše nebudou téměř nikomu povědomé.

7.4.3 Oblíbené seriály

Tato síť, s vazbami vytvořenými na základě shodnosti v oblíbených seriálech, je větší, než výše uvedená. Má 8966 hran pro 230 vrcholů. Je to způsobeno tím, že seriálů je méně než filmů a lidé tedy mají častěji mezi oblíbenými stejné. Grafické znázornění sítě je na obrázku 22. I tentokát byly do tabulky 7 vypsány hodnoty zkoumaných centralit pro prvních deset uživatelů s nejvyšší eigenvector centralitou a deset s nejmenší eigenvector centralitou, s tím, že uživatelé s nulovou degree centralitou byli odebráni. Výsledné hodnoty eigenvector centralit a Katz centralit jsou opět velmi obdobné. Pouze u spodní desítky uživatelů se projevil parametr β a Katz centrality jsou tedy o něco vyšší. Hodnoty PageRanku se také

Uživatel	Degree	Eigenvector	Katz	PageRank
Yoda666	1	1	1	1
silentfall	0,9273	0,8227	0,8283	0,797
Savanah2	0,7636	0,7839	0,7785	0,6827
bobrux	0,8545	0,7756	0,7884	0,8103
mark.l	0,7636	0,7441	0,7478	0,7161
Str3jd4	0,9091	0,7426	0,7625	0,9035
Masrii	0,7636	0,7035	0,7059	0,6391
epicek	0,4727	0,6629	0,6553	0,5147
komediant	0,8182	0,6605	0,675	0,7374
hanzi_111	0,7273	0,6472	0,6591	0,6552
Peregral	0,0545	0,0017	0,0525	0,1842
slatkac	0,0545	0,0017	0,0522	0,2504
Jasmeenka	0,0364	0,0015	0,0509	0,1549
Tomajko	0,0182	0,0014	0,0493	0,0838
Dollphin	0,0182	0,0011	0,0492	0,0739
zdeminem	0,0182	0,0002	0,0485	0,0954
Lucky74	0,0182	0	0,0483	0,1183
rudelino	0,0364	0	0,0495	0,3156
Osserk	0,0364	0	0,0495	0,3156
stanleyb7	0,0364	0	0,0495	0,3156

Tabulka 6: Tabulka centralit uživatelů z města Ostravy - oblíbené akční filmy



Obrázek 22: Graf sítě uživatelů z města Ostravy - hrany dle oblíbených seriálů

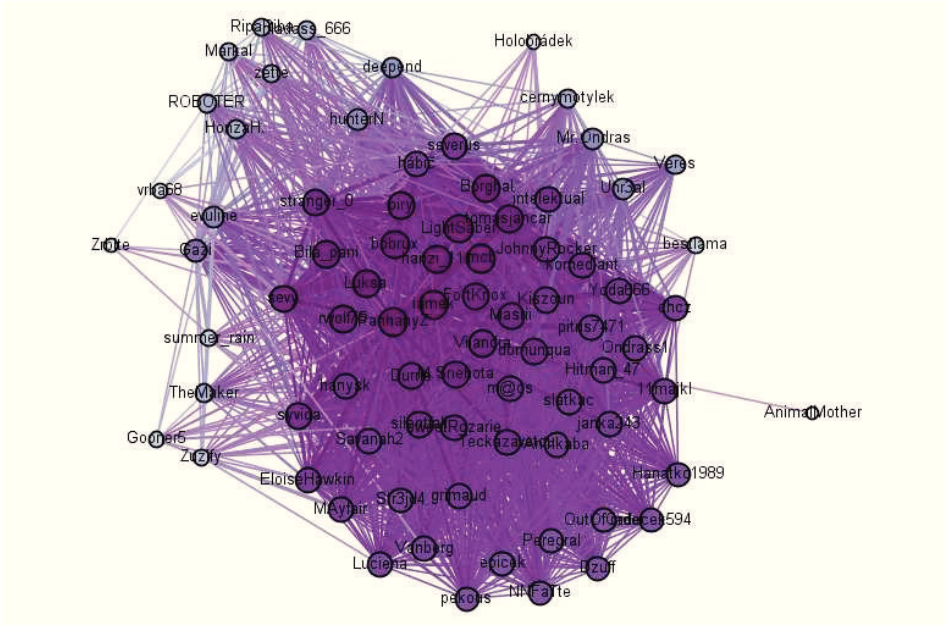
Uživatel	Degree	Eigenvector	Katz	PageRank
hitcher	0,9573	1	1	1
Teckazavetou	1	0,9605	0,9657	0,9997
epicek	0,9268	0,9233	0,9273	0,9264
hanzi_111	0,9695	0,9212	0,9259	0,9345
komediant	0,9817	0,9201	0,9249	0,9338
iamek	0,9268	0,9022	0,9064	0,9011
Kovi34	0,9512	0,8847	0,8924	0,9058
summer_rain	0,9573	0,8841	0,8905	0,9004
Cinka	0,9085	0,853	0,8585	0,8356
TheStapler	0,8963	0,8423	0,8486	0,8229
gianelka	0,1463	0,0563	0,1104	0,1327
mil@n.fanatik	0,1646	0,0554	0,1095	0,1413
RadimT	0,1097	0,0353	0,0905	0,1116
vrba68	0,0854	0,0307	0,0871	0,1250
11majkl	0,0609	0,0136	0,0702	0,0977
Jump3R	0,0487	0,0131	0,0695	0,0929
ssteav	0,0366	0,0129	0,0687	0,0839
zette	0,0366	0,0052	0,0623	0,0918
freemen	0,0122	0,0025	0,0589	0,0748
ROBOTER	0,0061	0,0014	0,0577	0,0722

Tabulka 7: Tabulka centralit uživatelů z města Ostravy - oblíbené seriály

příliš neliší. Uživatelé *hitcher* a *Teckazavetou* získali téměř identické ohodnocení, přestože *Teckazavetou* má relativně významně vyšší degree centralitu a *hitcher* naopak eigenvector centralitu. Zdá se tedy, že PageRank je nelépe vypovídající, jakýsi "zlatý střed". Podle konkrétního výběru seriálů těchto uživatelů je velmi obtížné určit kdo z nich má mainstreamovější vkus. Oba mají ve výběru velice známé a oblíbené seriály a bez výpočtu centralit je nemožné říci, kdo z nich je typičtějším divákem.

7.4.4 Oblíbení skladatelé

Jako poslední se podíváme na naši populaci uživatelů z Ostravy z pohledu jejich oblíbených skladatelů. Tato síť je podle očekávání nejmenší. Z 230-ti uživatelů má alespoň jednu vazbu pouze 76 uživatelů. V grafu znázorněném na obrázku 23 je celkem 2034 hran. Své oblíbené uživatele má vypsáno nejméně uživatelů, ve výsledné síti je tedy na naše poměry málo hran. Vypočtené hodnoty zkoumaných centralit uživatelů, vybraných podle stejných kritérií jako v předchozích případech, jsou v tabulce 8. Opět zde můžeme zahlédnout jistá známá jména. Například uživatel *hanzi_111* byl představitelem uživatele s typickým vkusem už u oblíbených seriálů. Máme tedy další případ poukazující na to, že typický vkus uživatelů se nevztahuje pouze na jednu jedinou oblast. Zajímavý je uživatel *PanhanyZ*, který má nejvyšší degree centralitu, ale v pořadí dle eigenvector centrality ob-



Obrázek 23: Graf sítě uživatelů z města Ostravy - hrany dle oblíbených skladatelů

sadil až páté místo. Přestože má jej reprezentující vrchol nejvíce hran, jedná se očividně o hrany se slabými váhami. Hodnota PageRanku je tradičně mezi eigenvector a degree centralitami. Podle PageRanku je uživatel *PanhanyZ* na třetím místě. Opět bychom mohli říci, že PageRank dává nejuvěřitelnější výsledky a zřejmě by bylo nejvýhodnější se řídit dle něj. Ale ani jedna ze zkoumaných centralit nedává vyloženě neadekvátní výsledky a řídit by se dalo podle všech. Záleží vždy na zvážení experimentátora.

Uživatel	Degree	Eigenvector	Katz	PageRank
mcb	0,9729	1	1	0,9780
hanzi_111	0,9729	0,9923	0,9954	1
FortKnox	0,9054	0,9743	0,9744	0,9441
LightSaber	0,9865	0,9566	0,9603	0,9517
PanhanyZ	1	0,9282	0,9348	0,9474
iamek	0,9865	0,9208	0,9282	0,9352
Masrii	0,8784	0,9066	0,9117	0,8736
Vilandra	0,9054	0,8982	0,9035	0,9284
M.Snehota	0,9054	0,8762	0,8834	0,8549
Durrie	0,8784	0,8388	0,8478	0,8136
Markal	0,3378	0,1292	0,1881	0,2159
RipaRiba	0,3378	0,1292	0,1881	0,2159
summer_rain	0,2432	0,1236	0,1803	0,1808
bestlama	0,2297	0,1183	0,1754	0,1741
Zuzify	0,2568	0,1115	0,1716	0,1978
Gooner5	0,2027	0,0785	0,1399	0,1595
Holobrádek	0,1216	0,0704	0,1301	0,1266
vrba68	0,1487	0,0539	0,1183	0,1508
Zrbite	0,0811	0,0309	0,0949	0,1106
AnimalMother	0,0135	0,0082	0,0729	0,0824

Tabulka 8: Tabulka centralit uživatelů z města Ostravy - oblíbení skladatelé

8 Závěr

Práce byla zaměřena na centrality odvozené od degree centrality v sociálních sítích. Byly podrobně popsány jak samotná degree centralita, tak eigenvector centralita, Katz centralita a PageRank. Teoretické poznatky byly aplikovány do praxe experimentováním nad zvolenou sociální sítí. Výpočty těchto centralit byly provedeny na vybraných podsítích sociální sítě Česko-Slovenské filmové databáze. Pro zisk dat z tohoto internetového serveru byl napsán vlastní program, který portál systematicky procházel a stahoval potřebné informace. Získaná data byla převedena do podoby grafů, reprezentujících sítě. S těmito sítěmi byly prováděny experimenty, kdy byly vlastním programem vypočteny výše uvedené centrality. Počítalo se na neorientovaných ohodnocených grafech, kdy vazby a jejich váhy byly vytvořeny na základě rozličných kritérií. Výsledné hodnoty centralit byly porovnány a na jejich základě jsme získali zajímavé informace týkající se vkusu uživatelů. Různými způsoby byli nalezeni uživatelé s nejtypičtějším vkusem, podle jejichž vkusu je pro většinu populace nejrelevantnější se řídit. Takovíto uživatelé byli nalezeni v různých sítích a podle různých kritérií. Díky těmto experimentům jsme také mohli porovnat zkoumané centrality a zdůvodnit případné odlišnosti v hodnotách daných centralit u stejných uživatelů. Byly tedy získány nejen praktické poznatky o centralitách odvozených od degree centrality, ale také užitečné informace o zkoumané sociální síti. Bylo například zjištěno, že uživatelé, kteří mají mainstreamový vkus ve filmech, ho mají většinou také v jiných oblastech jako oblíbené seriály či režiséři. Vytvořené aplikace pro zisk a zpracování dat mohou být použity pro další experimenty na odlišných sítích a sloužit pro průzkum konkrétních uživatelských skupin a konkrétních oblastí zájmů. Při tvorbě této práce bylo získáno mnoho teoretických i praktických poznatků z oblasti expertýzy sociálních sítí, ale i jiných informatických dovedností. Autor práce se naučil automatizovaně procházet internetový portál a získavat z něj potřebná data, procvičil si operace s SQL databází, vymyslel algorimy pro vytváření vazeb mezi uživateli na základě shody jejich hodnocení, prozkoumal možnosti ukládání grafových reprezentací sítí do textových souborů a jednu z těchto možností aplikoval, po nastudování teorie centralit odvozených od degree centrality, napsal algoritmy tyto centrality vypočítavající. Pro kontrolu výsledků se naučil pracovat s Pythonovou knihovou NetworkX a pro jejich grafickou reprezentaci s programem Gephi. Na základě získaných informací poté vyvodil závěry o specifících zkoumané sociální sítě.

9 Reference

[1] NEWMAN, M., *Networks: an introduction.* New York: Oxford University Press, 2010, xi, 772 p.

[2] FREEMAN, L. C., *Centrality in social networks: Conceptual clarification*, 1978, roč. 1, č. 3.

[3] Social networks: prestige, centrality, and influence. RUSINOWSKA, Agnieszka, Rudolf BERGHAMMER, Harrie C M De de SWART a Michel GRABISCH., *Relational and algebraic methods in computer science: 12th international conference, RAMICS 2011, Rotterdam, The Netherlands, May 30 - June 3, 2011. proceedings.*, 1st ed. New York: Springer, 2011, s. 22-39.

[4] TSVETOVAT, Maksim a Alexander KOUZNETSOV., *Social Network Analysis for Startups.* O'Reilly Media, 2011.

[5] BONACICH, Phillip., *The American Journal of Sociology: Power and centrality: a family of measures.* 1987, s. 1170-1182.

[6] HANNEMAN, Robert a Mark RIDDLE., *Introduction to Social Network Methods.* Riverside: University of California, 2005.

[7] KWAIT, Jennafer, Thomas VALENTE a David CELENTANO, *Interorganization Relationships Among HIV/AIDS Service Organizations in Baltimore: A Network Analysis. [online].* 2001, s. 468-487 [cit. 2013-03-22]. Dostupné z: http://www-hsc.usc.edu/∼tvalente/Publications/kwait_Valente_JUH.pdf

[8] KLEINBERG, Jon M., *Authoritative sources in a hyperlinked environment. Journal of the ACM.* roč. 46, č. 5, s. 604-632.

[9] XING, Wenpu a Ali GHORBANI., *Weighted pagerank algorithm.* In Second Annual Conference on Communication Networks and Services Research CNSR'04, pages 305–314, Fredericton, N.B., Canada, 2004.

[10] ČADA, R., T. KAISER a Z. RYJÁČEK., *Diskrétní matematika [online].* Plzeň: Západočeská univerzita v Plzni, 2004.

[11] BĚLOHLÁVEK, Radim a Vilém VYCHODIL., *Diskrétní matematika pro informatiky II.* Olomouc, 2006.

[12] KOVÁR, Martin., *Diskrétní matematika.* 2003.

[13] ADAR, Eytan., *The Graph Exploration System [online].* 2007 [cit. 2013-03-23]. Dostupné z: http://graphexploration.cond.org/

[14] *Gephi. GDF Format [online].* 2012 [cit. 2013-03-23]. Dostupné z: https://gephi.org/users/supported-graph-formats/gdf-format/

[15] *Gephi - PageRank. [online].* 2011 [cit. 2013-05-02]. Dostupné z: http://massapi.com/source/gephi-0.8-alpha.sources/StatisticsPlugin/src/org/gephi/statistics/plugin/PageRank.java.html

[16] *ALGLIB [online].* 2013 [cit. 2013-03-23]. Dostupné z: http://www.alglib.net/

[17] *NetworkX [online].* 2013 [cit. 2013-03-23]. Dostupné z: http://networkx.github.com/

[18] *NetworkX - katz.py. Github.com [online].* 2013 [cit. 2013-03-23]. Dostupné z: https://github.com/ComplexSystemTelecomSudParis/networkx/blob/master/networkx/algorithms/centrality/katz.py

[19] *NetworkX - pagerank_alg.py. Github.com [online].* 2013 [cit. 2013-03-23]. Dostupné z: https://github.com/ComplexSystemTelecomSudParis/networkx/blob/master/networkx/algorithms/link_analysis/pagerank_alg.py

A Obsah přiloženého DVD

DVD:/

- APLIKACE
  - **CSFD DataGainer** - projekt Visual studia aplikace pro stahování dat
  - **Centrality** - projekt Visual studia aplikace pro výpočty centralit
- GRAFY SÍTÍ - vytvořené grafy zkoumaných sítí
- VYPRACOVÁNÍ - diplomová práce ve formátu PDF